KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Extraction of Surface Characteristics and Lighting in 3D Reconstruction from Uncalibrated Images

Stamatios Georgoulis

Supervisor:
Prof. dr. ir. L. Van Gool

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

August 2017

Extraction of Surface Characteristics and Lighting in 3D Reconstruction from Uncalibrated Images

Stamatios GEORGOULIS

Examination committee:
Prof. dr. A. Bultheel , chair
Prof. dr. ir. L. Van Gool, supervisor
Prof. dr. ir. T. Tuytelaars
Prof. dr. ir. D. Vandermeulen
Prof. dr. ir. P. Hanselaer

Prof. dr. O. Sorkine-Hornung
(ETH Zurich)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

August 2017

Uitgegeven in eigen beheer, Stamatios Georgoulis, Kasteelpark Arenberg 10 - box 2441, 3001 Leuven (Belgium)

Acknowledgements

Finally, the time to fill these lines with gratitude words has come for me too. Yet, I can not help but feel a bittersweet taste in my mouth, as on the one hand my long PhD journey is about to end, but on the other hand during this journey I was lucky enough to be surrounded by extraordinary people who helped me, and motivated me to achieve this goal.

I can not think of a better person to start this section with than my supervisor, Prof. Luc Van Gool. Thank you Luc, for honoring me with the privilege to be one of your PhD students. Thank you, for giving me the opportunity to pursue my PhD dream, and the freedom to carve my own research path. Thank you, for never doubting my skills even when things did not go according to the plan, and the rejected submissions were much more than the accepted ones. Finally, thank you, for letting me to continue working with you as a post-doctoral researcher, so that I can squeeze every ounce of extra knowledge out of you.

I would also like to thank Prof. Tinne Tuytelaars, Prof. Mario Fritz, Prof. Tobias Ritschel, and Dr. Konstantinos Rematas, without whom the second half of this thesis could not be realized. Thank you Tinne, for acting as a co-supervisor the last couple of years, for having so many meetings with me, and for allowing me to learn so much from your research ideas. Thank you Mario, for allowing me to work with you, for sharing so many insightful discussions with me, and for adding your touch of perfectionism in every submitted text. Thank you Tobias, for introducing me to the computer graphics topics that became big part of this thesis, for being there whenever I needed you (especially non-office hours), for doing work that usually only PhD students do (I will never forget the manual annotations and videos), and for questioning every aspect of our submissions with a critical mind which helped me be prepared in advance for the challenging rebuttals. Thank you Kostas, for letting me join this wonderful research team that you shaped during your PhD, for sharing knowledge and code with me, for always being there on Google hangouts to cater my crazy demands, and for waking up at 7 o'clock in the morning just to

have meetings with us.

A big thank you goes to all my other co-authors, Dr. Marc Proesmans, Vincent Vanweddingen, Davy Neven, and Bert De Brabandere. Thank you Marc, for putting up with all my stupid questions when I first started my PhD, for sharing with me so much of your never-ending knowledge on 3D reconstruction, photogrammetry, rendering, and so on, and for instantaneously solving so many bugs on my code. Thank you Vincent (the Second) for sticking with me from day 1 in ESAT, for being a great office mate and friend, for doing all the labor-intensive scanning work whenever we had a real experiment, for rendering a tremendous amount of training data, and for so much more. Thank you Davy and Bert (the Young) for welcoming me in your Toyota sub-team, and for sharing your research knowledge and code with me.

This thesis would not reach its current state without the insights and helpful comments of my examination committee. Thank you Prof. Tinne Tuytelaars (again), Prof. Dirk Vandermeulen, Prof. Peter Hanselaer, Prof. Olga Sorkine-Hornung, and Prof. Adhemar Bultheel.

Of course all work and no play is boring. Thankfully, PSI-VISICS is a wonderful working environment. The acknowledgements for this must go to the rest of my (former and current) colleagues in Leuven. Many times I was saved by their indispensable assistance: be it a theoretical discussion with Angelo, an insightful paper recommendation from David, or an implementation trick from Bert (the Guru); with each bit I got a step closer to my goal. Amir, it was a pleasure sharing an office with you and your crazy music. To my teaching assistant mate Vincent (the First), your multi-color drawings in the blackboard shall be missed. There are many others I should thank (vaguely in order of appearance): Basura, Chris, Frank, Hakan, Honza, Jose, Wim, Rosalia, Xu, Jay, Stratis, Ali, Gina, Xuanli, Rahaf, Yu-Hui, Klaas, Thomas, Maxim, Liqian, Amal... the list goes on. Thank you everyone, for our initial lunches in Alma, deadline nights, birthday cakes, coke breaks, 'Game of Thrones' discussions, board games, bike trips, barbecues, beach volleyball, and beers on Oude Markt.

A big thank you goes to all of my relatives and friends; the ones in Greece, the ones in Leuven, and the ones that are now all over the world pursuing their dreams. Thank you relatives and friends in Thessaloniki: yiayia Eleni, theia Despoina, theia Toula, theios Goulis, Vaso, Eleni, theia Mairi, theios Antonis, Telis, Panagiotis, and Maria, Ria. Thank you relatives and friends in Sifnos: theia Maria, theios Thomas, Eleni, Maroula, theia Rita, Stamatis, Lora, theia Flora, Eleni, Katerina, and Anthoula, Maria. Thank you Athenians: Evgenia, Eleonora & Pavlos, Myrto. Thank you friends in Leuven: Niki & Antonis, Nadia & Dimitris, Danai & Stefanos, Argyro & Odysseas, Polina & Nikos, Rena & Manu, Eirini & Kristof, Silvia & Jonas, Carmen & Jasper, Spyros, George, Sofia,

Mitsos, Rodopoulos, Orestis, Maria, Magdalena, Nitesh. And so many more that my small brain may miss right now. You may not see your contribution right away but I can assure you that a happy life out-of-the-office is an essential step for PhD completion.

I could never thank enough the people from Thessaloniki, team "Epione", that introduced me to the vast academic world. Thank you Prof. Leontis (Hadjileontiadis), for being a research role model and making me realize that I want to pursue a PhD too. Thank you "Epione" gang, Stef, Dim, Kostas (the 'Ela rei'), Vag, Paris, for the unforgeable hours we spend on the notorious 6^{th} floor, trying to implement what seemed impossible. A special thank you to Dim and Stef for putting up with my psychological ups-and-downs for the whole PhD duration and for always encouraging me to continue no matter what. This PhD could not have been completed without you guys.

I have reserved this last part of the acknowledgements section for some special people in my life.

This thesis is dedicated to my family, mama Antigoni, mpampas Antonis, and bro Aris, for all the reasons that are impossible to describe in words. I know that I usually do not openly express my feelings, but I want you to know that without your love and unconditional support it would not be possible for me to achieve anything in my life. Thank you for letting me take all the wrong decisions in this journey, and thank you for making my problems yours. Last but not least, thank you Elena for being my brightest sunshine in the cloudy Leuven. Thank you all for everything! I love you all very much!

Stamatios Georgoulis
Leuven, August 2017

Abstract

A cliche expression is that "an image is worth a thousand words". That is to say, when humans are given a single image (or a sequence of images) they are very accurate in "extracting" information about the surface geometry and characteristics or the incident lighting just by observing the physical interactions between the surfaces and the light sources in the scene. When it comes to computers, however, it is still questionable to what extent they can achieve similar results. If we carefully examine the image formation process, *i.e.* light emitted from a source is reflected, absorbed or transmitted between surfaces before entering the camera's aperture, we observe that this procedure naturally imprints information about the surfaces and light sources in the scene. As such, there is merit in training computers on how to undo the image formation process, as humans unconsciously do.

The latter practically means that we want to decompose an image or a sequence of images into their intrinsic 3D geometry, surface reflectance and incident illumination, so that these individual components if modified and re-synthesized result in a photo-realistic rendering of the original scene. Unsurprisingly, recovering these hidden components from sheer images alone is an important problem with many applications in computer vision, computer graphics and machine learning tasks. In this thesis, we address this problem and try to infer information about the surface (*i.e.* the geometry and reflectance properties) as well as the light sources (*i.e.* the environmental illumination) by carefully observing the pixel values of the input image(s).

However, the decomposition of a scene into its intrinsic components given such small amount of information as input is a very difficult and under-constrained task, as the same visual result might be due to many different combinations of intrinsic components. Furthermore, real-world scenes exhibit complex reflectance behavior and incident lighting in contrast to assumptions made in literature, such as the surface reflectance is limited to being purely diffuse or can be fully described by parametric models, and the scene's illumination only comes from

infinitely-distant point lights. Even most recent approaches, that have tried to use less strict assumptions on reflectance and illumination, still require the capture of high-dynamic range images (*i.e.* taking many pictures under different exposures) and the use of dedicated hardware setups, making these approaches rather impractical and inaccessible to casual users.

Motivated by these observations, in this thesis we build a method for the inference of complex, real-world reflectance and natural illumination and the refining of geometry from a single image or a sequence of images and rough or exact initial geometry. We develop a framework of tools, including surface reflectance capture and inference, illumination estimation, and geometry refinement techniques to efficiently solve the problem of decomposing the geometric and radiometric information of the scene imprinted onto the image(s). We demonstrate the effectiveness of our framework on a large set of synthetic and real data and give in-depth quantitative and qualitative evaluation of our method as well as comparisons with state-of-the-art approaches.

Beknopte samenvatting

Het is een cliché dat ëen beeld meer zegt dan duizend woorden". Als mensen één enkele afbeelding (of een reeks van afbeeldingen) te zien krijgen, kunnen ze heel accuraat informatie over de geometrie en eigenschappen van het invallend licht achterhalen door enkel de fysische interactie tussen de oppervlakken en de lichtbronnen in de scene te observeren. Het is echter maar de vraag of computers gelijkaardige resultaten kunnen halen. Als we zorgvuldig het beeldvormingsproces bekijken, *ttz.* licht uitgezonden door een bron reflecteert, wordt geabsorbeerd of breekt tussen oppervlakken alvorens de lens van de camera te bereiken, kunnen we begrijpen dat dit proces op een natuurlijke wijze informatie over de oppervlakken en lichtbronnen incorporeert. Daarom is het opportuun computers te trainen dit beeldvormingsproces ongedaan te maken, zoals mensen ook onbewust doen.

Het laatste betekent praktisch dat we een afbeelding of een sequentie van afbeeldingen willen uiteenrafelen in de intrinsieke 3D geometrie, oppervlaktereflectantie en invallende illuminantie zodat er indien deze componenten worden aangepast en geresynthetiseerd, een fotorealistische rendering van de originele scene ontstaat. Het hoeft niet te verbazen dat het achterhalen van deze verborgen componenten louter uit afbeeldingen een belangrijk probleem met veel toepassingen is in computervisie, computergrafieken en machinaal leren. In deze thesis addresseren we dit probleem en proberen we informatie af te leiden over het oppervlak (*ttz.* de geometrie en reflectie-eigenschappen) alsook de lichtbronnen (*ttz.* het omgevingslicht) door zorgvuldig te kijken naar de pixelwaardes van de invoerafbeeldingen.

De decompositie van een scene in haar intrinsieke componenten, als maar een kleine hoeveelheid informatie beschikbaar is, is een heel moeilijke en onderbegrensde taak. Dezelfde visuele uitkomst kan immers bekomen worden door heel veel verschillende combinaties van de intrinsieke componenten. Daarenboven zijn de reflectie-eigenschappen en het invallend licht in de echte wereld complexer te achterhalen dan de veronderstellingen die in de literatuur

gebruikt worden (zoals bijvoorbeeld dat de oppervlaktereflectantie enkel diffuus of volledig door parametrische modellen kan beschreven worden en dat de lichtbronnen in de scene puntbronnen zijn die oneindig ver staan). Nog recentere aanpakken, die minder restrictieve veronderstellingen hebben, stellen nog altijd hoge eisen aan de invoerafbeeldingen (hoog dynamisch bereik, door het nemen van meerdere foto's na elkaar met verschillende sluitertijden) of moeten gebruik maken van gespecialiseerde hardware setups, wat deze methoden onpraktisch en onbereikbaar maakt voor reguliere gebruikers.

Met deze observaties in gedachten, construeren we in deze thesis een methode voor de afleiding van complexe, echte-wereld reflectantie en natuurlijke belichting en het verfijnen van de geometrie, uitgaand van een enkele afbeelding of een reeks afbeeldingen en ruwe of exacte initiële geometrie. We ontwikkelen een framework van hulpmiddelen, *oa.* methodes om oppervlakereflectantie op te meten en af te leidingen, het schatten van de belichting en geometrieverfijningstechnieken om efficient het probleem van het decomposeren van de geometrie en de radiometrische informatie van de scene die ingebakken zit in de afbeelding(en) op te lossen. We demonstreren de effectiviteit van ons framework op een grote set van synthetische en echte data en geven een grondige quantitatieve en kwalitatieve evaluatie van onze methode alsook vergelijkingen met state-of-the-art methodes.

Abbreviations

?D	? Dimensional
ADM	Alternating Direction Method
AL	Augmented Lagrangian
AR	Augmented Reality
ARC3D	Automatic Reconstruction Cloud
BRDF	Bidirectional Reflectance Distribution Function
CAD	Computer-Aided Design
CAM	Computer-Aided Manufacturing
CCD	Couple-Charged Device
CG	Conjugate Gradient
CIE	Commission Internationale de l'Eclairage
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Convolutional Neural Network
DoG	Difference-of-Gaussians
DS-GPLVM	Discriminative Shared GPLVM
DSLR	Digital Single-Lens Reflex (camera)
DSSIM	structural DiSSIMilarity (metric)
EXIF	EXchangeable Image File
FOV	Field-Of-View
GP	Gaussian Process
GPLVM	Gaussian Process Latent Variable Model
GPU	Graphics Processing Unit
HDR	High-Dynamic Range
IBP	Independent Back-Projections
KRR	Kernel Ridge Regression
LAB	A color space defined by the CIE
LDP	Location Determination Problem
LDR	Low-Dynamic Range
LED	Light Emitting Diode
LOO	Leave-One-Out

LRMSE	Logged RMSE
MatConvNet	A MATLAB toolbox implementing CNNs
MERL	Mitsubishi Electric Research Laboratories
MSE	Mean Square Error (metric)
MvS	Multi-view Stereo
NN	Nearest Neighbor
PCA	Principal Component Analysis
PS	Photometric Stereo
PSR	Poisson Surface Reconstruction
RANSAC	RANdom SAmpling Consensus
RAW	Native file format generated by the camera
RBF	Radial Basis Function
RE	Rendering Equation
ReLU	Rectified Linear Unit
RGB	Red Green Blue (image)
RGB-D	Red Green Blue Depth (image)
RM	Reflectance Map
RMSE	Root MSE (metric)
SfM	Structure-from-Motion
SH	Spherical Harmonics
SIFT	Scale-Invariant Feature Transform
SL	Structured Light
SMASHINg	Specular MAterials on SHapes with complex IllumiNation
s-o-t-a	state-of-the-art
SURF	Speeded Up Robust Features
VR	Virtual Reality

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Given a single image or a sequence of images depicting an object, humans are very accurate in "estimating" information about the object's surface characteristics or the environmental lighting just by observing how light in the visible spectrum is reflected by the object under the environmental lighting. This human visual perception ability persists even when the particular object has never been seen before; just a few reflectance observations are enough to facilitate the imaginary decomposition process that takes place in our brains. For example, when we see a picture of an athlete being nominated a silver metal in the Olympic Games, we can immediately understand that the medal has a disk-like shape, is probably made of metal due to an almost mirror-like characteristic reflection, and that the sun is above the athlete because there is a very strong reflection on the upper part of the medal and the nomination ceremony takes place outdoors.

The question that naturally raises is whether nowadays computers can achieve similar, if not better, results starting from the same image or sequence of images. To answer this question we should first go back to how images are formed when using a camera. Camera-captured images are the result of physical interactions between surfaces and light sources. Specifically, light begins by emission from a source and is reflected, absorbed, or transmitted between surfaces in the scene, before potentially entering the camera and interacting with the film in analog cameras or digital imaging sensor (*e.g.* Couple-Charged Device (CCD), Complementary Metal-Oxide-Semiconductor (CMOS)) in digital cameras. In this thesis we investigate the latter case. Consequently, the camera filters the light by wavelength range, such that the three separate filtered intensities (red, green and blue) include information about the color of light. After some processing steps (*e.g.* de-noising, white balancing, gamma correction) the - RGB

- image is finally formed.

During this image formation process the amount of light and the path the light follows from emission from a source till interaction with the imaging sensor of the camera depends on the geometry of the scene as well as the radiometric properties of each surface in the scene. Specifically, when a ray of incoming light is reflected or transmitted it encapsulates information about the surface and the environmental lighting into the outgoing light ray. Moreover, the outgoing light ray will be attenuated in different spectra and sent along a unique distribution of directions depending on the surface characteristics of the object. As such, the image formation process imprints information about the surface and light sources in the scene into the outgoing light ray, allowing us to infer information about the surface (*i.e.* the geometry and reflectance properties of the object) as well as the light sources (*i.e.* the environmental illumination) by carefully observing the pixel values of the image.

Motivated by these observations, in this thesis we build a method for the inference of complex, real-world reflectance and natural illumination and the refining of geometry from a single image or a sequence of images and rough or exact initial geometry. We develop a framework of tools, including surface reflectance capture and inference, illumination estimation, and geometry refinement techniques to efficiently solve the problem of decomposing the radiometric information of the scene imprinted onto the image(s). Unlike previous approaches that tried to tackle this problem using sophisticated hardware setups that have proven to be complicated, expensive and most importantly inaccessible to casual users, in this thesis we limit ourselves to the use of readily available consumer equipment. We demonstrate the effectiveness of our framework on a large set of synthetic and real data and give in-depth quantitative and qualitative evaluation of our method as well as comparison with state-of-the-art (s-o-t-a) approaches.

In Section 1.1 we present the motivation that drives our work. Then, in Section 1.2 we present the main tasks that this thesis is focusing on. Section 1.3 poses the research questions that will be explored throughout the thesis. Finally, in Section 1.4 we give a brief overview of the thesis and present our main contributions to the topic.

1.1 Motivation

In recent years the problem of recovering geometric and radiometric information from plain images has received considerable attention. Shape-from-shading [66], a method for learning the geometry of an object from an image under strict

assumptions, has pioneered the way for recovering geometric and radiometric properties from a set of images. To overcome the strict assumptions posed by [66], Barron and Malik [8] proposed a method for inferring spatially-varying reflectance, spatially-varying illumination, and geometry from RGB-D images, and Zickler *et al.* [171] developed an approach to measure spatially-varying reflectance from a small set of images. However, limiting the reflectance and illumination models, in the sense that either the surface reflectance is assumed to be purely diffuse [8] or the scene's illumination only comes from an infinitely-distant point light [171], also limits the amount of information we can recover from the scene.

In the real world, reflectance behavior and incident lighting is complex. For example, real-world scenes have complex reflectance functions, often featuring off-specular peaks, retroreflection, and subsurface scattering effects, that can vary along a surface. Furthermore, objects are illuminated from arbitrary directions in the scene - not only by single point light sources but by distant surfaces in the scene (*e.g.* the sky, the walls, the ground). Recently, Oxholm and Nishino [110] and Lombardi and Nishino [93] modeled complex real-world reflectance functions and natural illumination for recovering geometric and radiometric properties from a set of High-Dynamic Range (HDR) images. Although they rely on less strict assumptions about the surface reflectance and the illumination of the scene, they still require the capture of (multiple) HDR images (*i.e.* taking many pictures under different exposures) and the use of dedicated hardware setups, making their approaches rather impractical and inaccessible to casual users. Therefore, there is still merit in extracting geometric and radiometric information from plain Low-Dynamic Range (LDR) images - especially after the re-appearance of deep learning - as the latter opens new possibilities for a broad set of applications domains.

Being able to extract accurate geometry, in the form of a 3D shape, has important implications in a variety of areas. In industrial design and manufacturing, it allows for developing new product designs, taking the measurements of objects with complex geometry (from a small mechanical part to a turbine) or automating the workflow at manufacturing facilities. The resulting 3D models can nowadays be exported to a variety of CAD & CAM programs and from there gauged and modified to improve the product's design and performance or integrate it into a new production system. In healthcare, it is successfully used to produce quick and accurate scans of the body or body parts for making 3D printed implants. In science and education, universities, colleges and laboratories are embracing 3D shape extraction as a powerful tool that allows students and researchers to study artifacts in greater detail than ever before without risking damaging them. The world's leading museums also use 3D scanning technologies to digitize artifacts and create online galleries,

facilitating access to their collections for art specialists and academics no matter where they are based. In arts and design, it drives forward the movie industry and video games - many stunts and visual effects would be difficult or even impossible to bring off before the advent of 3D shape extraction.

Decoding the radiometric properties of the scene is also an important problem with many applications. On the one hand, the reflectance of an object can give us important clues about what the object is made of. For example, an object made of gold has a distinctive appearance: it is a golden-yellow color and it reflects light in an almost mirror-like way. If we could identify these qualities, we could recognize the material of that object, which in turn enables the development of new algorithms and new technologies. A mobile robot or autonomous automobile can use material recognition to determine whether the terrain is asphalt, grass, gravel, ice or snow in order to optimize mechanical control. An indoor mobile robot can distinguish among wood, tile, or carpet for cleaning tasks. The potential applications are limitless. On the other hand, the illumination of the scene can give us important information too. If we knew that a scene was brightly lit from a single direction with a particular spectral signature, we might conclude that the sun is illuminating the scene and that the scene, therefore, takes place outdoors. We could even take this a step further and draw conclusions about where in the world the picture was taken based on the angle of the sun and the current time.

The above factors drive the work presented in this thesis, where we investigate how to recover accurate geometric and plausible yet realistic radiometric information from a single LDR image or a sequence of LDR images.

1.2 Tasks of Interest

Computer vision aims at tackling a variety of problems, *e.g.* scene recognition, 3D reconstruction, image classification, object detection, video tracking, pose estimation, learning, to name a few. All of the aforementioned tasks follow a common pattern: they take as input an image or a sequence of images (*i.e.* video) and they output an *understanding* of the processed input. Depending on the task, this understanding can be a class label for image classification, a bounding box for object detection, a 3D shape for 3D reconstruction, etc. In computer graphics, however, given an understanding of the world we aim to synthesize an image. For example, the rendering process aims at generating a 2D image given an analytic description of a 3D scene. Thus, generally speaking one could see computer graphics tasks as the inverse procedure of computer vision tasks and vice versa. This thesis lies in the intersection of computer vision and computer

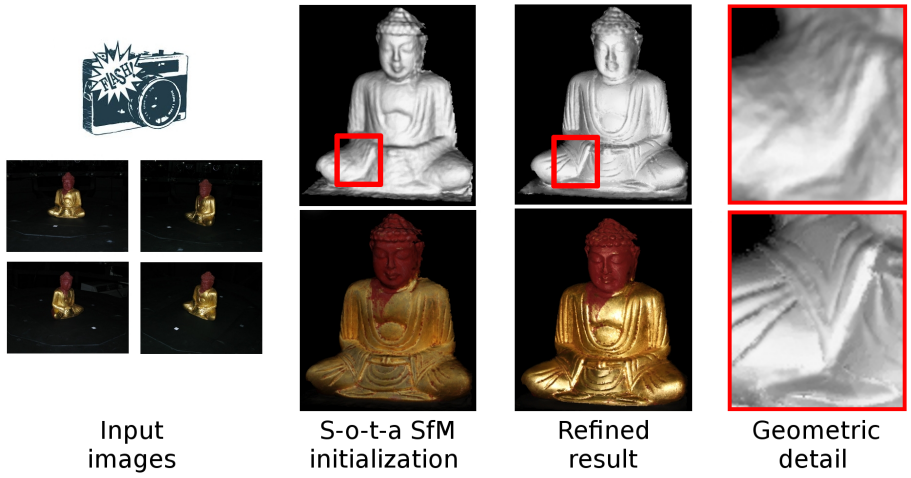


Figure 1.1: A primary goal of this thesis is the estimation of geometric and radiometric information from images. Left: The user should simply walk around the object of interest taking pictures under the illumination of the camera's flash. Middle-left: Using s-o-t-a uncalibrated SfM, we extract a rough initial shape rather smooth and a diffuse reflectance rendering. Middle-right: Using principles from PS, in this thesis we propose a method to arrive at a new estimate with refined geometry and photo-realistic reflectance. Right: Geometry detail between initial (top) and refined (bottom) mesh.

graphics and focuses on undoing the rendering process, *i.e.* decompose a scene into its intrinsic 3D shape, surface reflectance and environmental illumination, so that these components if edited and re-synthesized they result in a photo-realistic rendering of the original scene. In what follows, we describe in more detail the tasks associated with each of these three basic components: 3D shape, surface reflectance, environmental illumination.

1.2.1 Extracting 3D Shape from 2D Images

Multiview 3D reconstruction is the task of creating 3D models (*i.e.* output) from a set of 2D images (*i.e.* inputs). In more detail, the essence of an image is a projection from a 3D scene onto a 2D plane, during which process the depth is lost. The 3D point corresponding to a specific image point is constrained to be on the line of sight. From a single image, it is impossible to determine which point on this line corresponds to the image point. If two or more images are available, however, then the position of a 3D point can be found as the intersection of the

two projection rays. This process is referred to as triangulation. The key for this process is the relations between multiple views, which convey the information that corresponding sets of points must contain some structure, and that this structure is related to the poses and the calibration of the camera.

In recent decades, there is an important demand for 3D content for computer graphics, virtual reality and communication, triggering a change in emphasis for the requirements. Many existing systems for constructing 3D models are built around specialized hardware (*e.g.* stereo rigs) resulting in a high cost, which cannot satisfy the requirement of its new applications. This gap stimulates the use of digital imaging facilities, like a camera. Moore's law also tells us that more work can be done in software. Uncalibrated Structure-from-Motion (SfM) [117] is a great example of such a software-based technique for estimating 3D structures from 2D image sequences that may be coupled with local motion signals. It is based on the principle that, as humans perceive a lot of information about the 3D structure in their environment by moving through it, 3D information can be obtained from images sensed over time.

However, in this thesis we are interested in solving a variant of this problem: given a rough initial 3D shape of an object, obtained using SfM, we want to refine its geometry using principles from Photometric Stereo (PS) [159], such that the final outcome closely resembles the ground truth object both in terms of geometric and radiometric details. Fig. 1.1 gives a preview. With such a task we assume that a rough initial 3D shape is always known in advance, but in terms of applicability, it is more attractive since the recovered geometric and radiometric information allows for photo-realistic renderings as well as numerous editing options (*e.g.* the change of the object's reflectance properties or the manipulation of its 3D shape).

1.2.2 Measuring Surface Reflectance Properties

The reflectance properties of a surface are described by the Bidirectional Reflectance Distribution Function (BRDF) [106], which is a function that defines how light is reflected at an opaque surface. It is employed both in the optics of real-world light, in computer graphics algorithms, and in computer vision algorithms. The function takes an incoming light direction and outgoing direction (taken in a coordinate system where the surface normal lies along the z-axis), and returns the ratio of reflected radiance exiting along the outgoing direction to the irradiance incident on the surface from the incoming direction. Both incoming and outgoing directions are parameterized by azimuth and zenith angles, therefore the BRDF as a whole is a function of four variables.

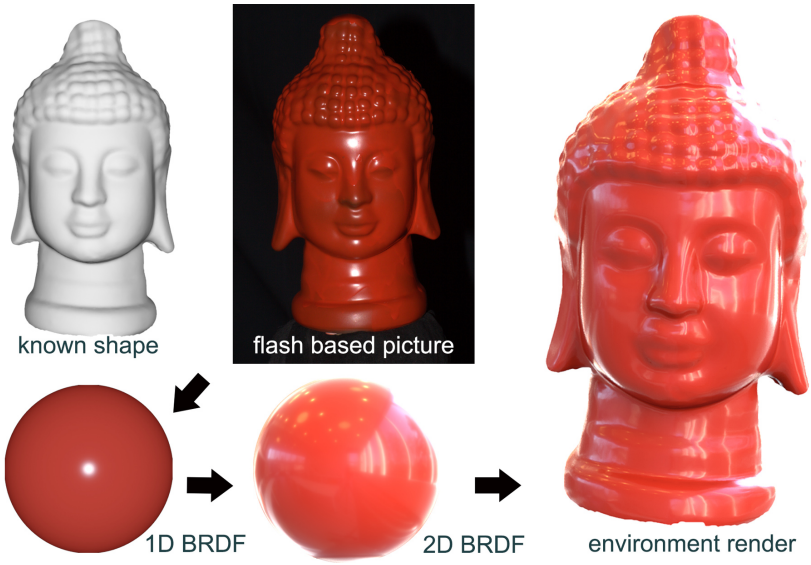


Figure 1.2: A secondary goal of this thesis is the inference of surface reflectance properties. Given a known shape (top-left) a low-dimensional BRDF (bottom-left) can be extracted from a single image (top-middle). In this thesis, we propose a method to infer the full BRDF (bottom-middle), in order to relight the object of interest under novel lighting conditions (right).

Traditionally, the task of measuring a BRDF is performed by sophisticated hardware devices called gonioreflectometers [27]. The device itself consists of a light source illuminating the material to be measured and a sensor that captures light reflected from that material. The light source should be able to illuminate and the sensor should be able to capture data from a hemisphere around the target. The hemispherical rotation dimensions of the sensor and light source are the four dimensions of the BRDF. The 'gonio' part of the word refers to the device's ability to measure at different angles. In general, gonioreflectometers are expensive and inaccessible to most researchers, let alone casual users. Furthermore, a dense sampling of an object's BRDF - usually only of a small planar patch - is a time-consuming process; for a sampling at an angular resolution of 1 degree more than 10^8 measurements are required [74].

To account for these limitations, in this thesis we start from a low-dimensional BRDF, which can easily be extracted from a single image of an object with known geometry, and try to infer the full high-dimensional BRDF, in order to relight the object under novel lighting conditions. Fig. 1.2 summarizes this procedure. Apart from drastically reducing the scanning time, our alternative

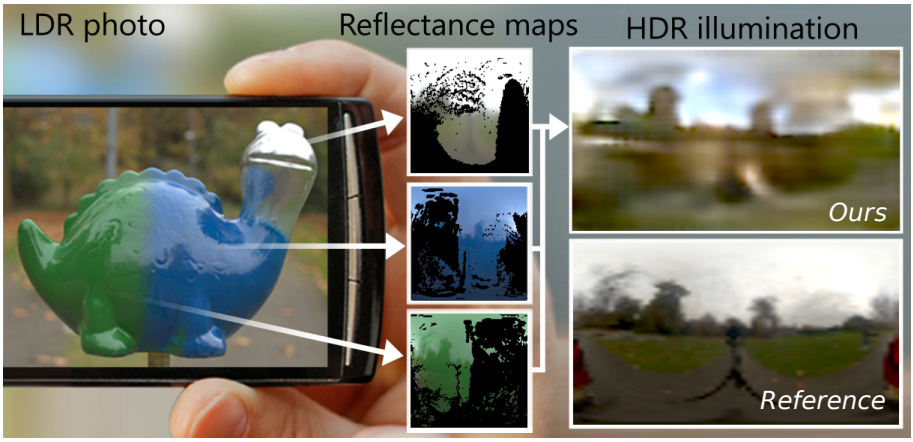


Figure 1.3: A final goal of our thesis is the estimation of environmental illumination in a more practical setup. To tackle this problem, in this thesis we propose a method to compute a HDR environment map (right-top) from an LDR photo of a multi-material object (left) by combining information from the different materials (middle) and the background.

approach for estimating the reflectance properties of a surface has interesting applications on material relighting and flash-based photography.

1.2.3 Capturing Environmental Illumination

Illumination (or lighting) is one of the most important aspects of object modeling. Like in the real world, every object looks different under different illumination conditions. In most cases, to properly model the illumination of a scene, one should take into account not only the light that comes directly from a light source (*direct illumination*), but also subsequent cases in which light rays from the same source are reflected by other surfaces in the scene, whether reflective or not (*indirect illumination*). In this thesis, we mainly focus on environmental (or natural) illumination, which mediates between direct and indirect illumination. Specifically, environmental illumination is not only limited to point light sources as it accounts for more natural lighting coming from every direction (*e.g.* a real scene with sky, windows, etc), but at the same time it does not include multiple bounces of the light rays on other surfaces in the scene.

Typically, environmental illumination is captured by taking an HDR image of a mirrored sphere, typically made of chrome steel [28]. In more detail, the mirrored sphere and the digital camera are placed on tripods about 2 meters

from each other and 1.5 meter off the ground in the scene of interest. The digital camera is zoomed until the sphere fills the frame, the focus is set so that the reflected light is sharp, and an HDR image series is acquired with shutter speeds varying from 1 to 1/10.000s, spaced one stop apart. To cover the scene with better sampling, a second HDR image series is acquired after having moved the camera 180° around to see the ball from the other side. Using special software the image series are converted into an omni-directional (360° panoramic) HDR image, namely an environment map or light probe [30]. Understandably, this is a time-consuming, labour-intensive and expensive process known only by experts and is also not an option for already existing footage or dynamic scenes.

In this thesis, we focus on drastically reducing the acquisition effort required for capturing environmental illumination maps. To do so, we use everyday objects - *i.e.* far-from-perfect-mirrors both in terms of shape and materials - to act as light probes (cf. the Dino in Fig. 1.3) and we compute an HDR illumination map from an LDR photo of a multi-material object (Dino) by combining information from the different materials and the background. Our simplified scanning procedure is particularly attractive in terms of applicability, allowing us to convert even images found on the internet into light probes, which in turn can be used to insert virtual objects in the same scene.

1.3 Research Questions

From the previous sections it is more than apparent that recovering geometric and radiometric information from plain images is a problem not yet fully explored. Especially when the input is a single image, where the given information is at best minimal, little progress has been made. Furthermore, there are several benefits from having a solid approach for estimating 3D shape, surface reflectance and natural illumination from images and the re-appearance of deep learning shows great potential towards this direction.

Given these observations, the objective of this thesis is to dive into methods for extracting 3D shape, inferring surface reflectance properties, and estimating environmental illumination from a single or a set of images. In particular, all these tasks can be summarized by the following research question:

Given a single image or a sequence of images depicting an object in a scene, can we recover its intrinsic 3D shape, surface reflectance and environmental illumination, so that these individual components if modified and re-synthesized result in a photo-realistic rendering of the original scene?

The components to be modified in this case can be any of the 3D shape, surface

reflectance or environmental illumination. For example, we can change the 3D shape of the object, edit its surface materials, place it in another scene (*i.e.* change the environmental illumination), insert another object in the same scene, etc, or any combination of the above. The possibilities are endless.

The decomposition of an object in a scene into its intrinsic components (3D shape, surface reflectance, environmental illumination) given such small amount of information as input (a single image or a small set of images) is a very difficult and under-constrained task, as the same visual result might be due to many different combinations of intrinsic object properties. For this reason and for the sake of presentation clarity, we split the main research question into smaller questions that address each of these components gradually, going from stricter to looser assumptions, before tackling this problem as a whole.

1. Can we extract 3D shape and surface reflectance from a small set of uncalibrated images and under which lighting conditions?
2. To what extent can we infer high-dimensional reflectance information from a single image?
3. How can we estimate the environmental illumination of a multi-material object given an image as the sole input?
4. Is it possible to decompose a single image into its intrinsic 3D shape, surface reflectance and environmental illumination and if so, what assumptions should be made to make this decomposition feasible?

The path shaped from answering these questions resulted in the contributions of this thesis, as presented in the following section.

1.4 Overview and Thesis Contributions

The work presented in this thesis focuses on the recovery of geometric and radiometric information from a single image or a small set of images and the necessary prerequisites to achieve this goal. The research conducted for this purpose has led to the publication of several papers and in this thesis for the sake of presentation clarity we present each paper to a separate chapter. In order to facilitate the reading flow, the related work is discussed in each chapter individually. Overall, the core of the thesis tries to answer the main research question posed in the previous chapter. In what follows, we provide a more detailed description of the contents of each chapter, followed by a reference to the respective paper where applicable.

In Chapter 2 we introduce the reader to the background knowledge that this thesis is built upon. In particular, we briefly describe the core research strands and datasets that were used throughout the thesis. By doing so, we aim at providing the reader with all the necessary information required in the following chapters of the thesis.

In Chapter 3 we investigate the use of simple flash-based photography to capture an object's 3D shape and reflectance characteristics at the same time. The presented method combines the principles of SfM, Multi-view Stereo (MvS) and PS, yet, we make sure not to use more than readily available consumer equipment, like a camera with flash or a smartphone. Starting from a rough low-resolution mesh generated from SfM and MvS, we apply a PS-based technique to refine both geometry and reflectance, where the latter is expressed in terms of data-driven BRDF representations. Compared to existing approaches our system fulfills three basic principles in order to bring shape and reflectance acquisition closer to non-expert users. First, only readily available consumer equipment is required for the scanning, like a flash-equipped DSLR camera or smartphone. Second, a small number of LDR images need to be recorded. Third, minimum to no intervention is required by the user (*e.g.* using calibration techniques, creating masks, etc). This is the first system providing to casual users the ability to extract both 3D shape and surface reflectance under such conditions. This chapter aims at answering the Research Question 1, and from our quantitative and qualitative analysis we essentially show that the answer is positive under the aforementioned lighting conditions. The contents of this work are based on the following publication[1]:

- S. Georgoulis, M. Proesmans and L. Van Gool, *Tackling Shapes and BRDFs Head-on*. Published in IEEE International Conference on 3D Vision (3DV) 2014.

Chapter 4 focuses on the problem of inferring higher order reflectance information starting from the minimal input of a single BRDF slice (*i.e.* a low-dimensional BRDF). In particular, we examine the prototypical case of a homogeneous sphere, lit by a head-on light source, which only holds information about less than 0.001% of the whole BRDF domain. We propose a novel method to infer the higher dimensional properties of the material's BRDF, based on the statistical distribution of known material characteristics observed in real-life samples. Unlike previous studies that consider either environmental lighting or sparse samples across the entire BRDF domain, in our case the coincidence of

[1]This work has been extended and included in the following paper: S. Georgoulis, V. Vanweddingen, M. Proesmans and L. Van Gool, *Shape and Reflectance Using a Camera with Flash*. Submitted in IEEE International Journal on Computer Vision (IJCV).

lighting and viewing directions only yields a small section of the BRDF space. This is a particularly difficult case compared to this considered in previous methods, because not only do we have very few samples but they are also very concentrated, so in our case inferring the rest of the BRDF is more a matter of extrapolation than interpolation. The proposed solution is general enough to deal with this issue, as well as to infer BRDFs of multiple dimensions. The content of this chapter addresses Research Question 2 and is based on the published paper:

- S. Georgoulis, V. Vanweddingen, M. Proesmans and L. Van Gool, *A Gaussian Process Latent Variable Model for BRDF Inference*. Published in IEEE International Conference on Computer Vision (ICCV) 2015.

In Chapter 5 we introduce our method for recovering natural illumination from a single LDR image. We propose a deep Convolutional Neural Network (CNN) that combines prior knowledge about the statistics of illumination and reflectance with an input that makes explicit use of two observations. First, images rarely show a single material, but rather multiple ones that all reflect the same illumination. Second, parts of the illumination are often directly observed in the background, without being affected by reflection. Our approach maps multiple partial LDR material observations represented as reflectance maps and a background image to a spherical High-Dynamic Range (HDR) illumination map. This is the first line of work that shows how to estimate HDR illumination from a single LDR image. This chapter answers the Research Question 3: the proposed method shows how both multi-material and using a background are essential to improve illumination estimations. The contents of this chapter are based on the following paper:

- S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars and L. Van Gool, *What Is Around The Camera*. Published in IEEE International Conference on Computer Vision (ICCV) 2017.

While the previous chapters deal with the estimation of one or two components that synthesize an image (3D shape, surface reflectance, environmental illumination), in Chapter 6 we investigate how to tackle all three at the same time. Specifically, we present a method that estimates reflectance and illumination information from a single image, where the input image depicts a single-material object of a given class with a specular material and under natural illumination. In contrast to earlier work, we follow a data-driven, learning-based approach and do not assume one or more components (shape, reflectance or illumination) to be known. To achieve this, we propose a two-step approach, where we

first estimate the object's reflectance map, and then further decompose the latter into reflectance and illumination. For the first step, we introduce a CNN that directly predicts a reflectance map from the input image itself, as well as an indirect scheme that uses additional supervision, first estimating surface orientation and afterwards inferring the reflectance map using a learning-based sparse data interpolation technique. For the second step, we suggest a CNN architecture to reconstruct both reflectance parameters (*i.e.* Phong parameters) and illumination (*i.e.* high-resolution spherical illumination maps) from the reflectance map. Our key contributions are summarized below. First, we propose the first deep learning formulation to infer reflectance maps from a 2D image and to further decompose them into material parameters and natural illumination. Second, we show new capabilities of CNN architectures, mapping from the image to the directional domain, performing learning-based sparse data interpolation as well as mapping from LDR to HDR data. Third, in order to train and evaluate our two-step approach, we provide new datasets that include large scale synthetic data to facilitate the training of deep learning models as well as real data to provide a realistic testing regime. The work in this chapter answers Research Question 4 and is included in the paper[2]:

- S. Georgoulis[3], K. Rematas[3], T. Ritschel, E. Gavves, M. Fritz, L. Van Gool and T. Tuytelaars, *Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning.* Published in IEEE Journal on Pattern Analysis and Machine Intelligence (PAMI) 2017.

Chapter 7 concludes the thesis. After summarizing the contributions of this thesis, we present the insights gained through answering the individual research questions posed in the previous section. Next, we analyze what are the limitations of this work. Finally, we discuss about possible future directions of our research.

[2]The first step of our two-step approach was originally published in the paper: K. Rematas, T. Ritschel, M. Fritz, E. Gavves and T. Tuytelaars, *Deep Reflectance Maps.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.

[3]S. Georgoulis and K. Rematas contributed equally to this work.

Chapter 2

Background

This chapter presents the foundations that this thesis was built upon. Since we deal with a variety of research problems in computer vision, computer graphics and machine learning, we find it fit to provide a small description of the basic concepts used throughout this thesis. We begin by presenting the image formation process in the camera. Then we proceed to traditional methods for extracting 3D shape that are necessary to either get a rough geometry that serves as an initialization to our approach or acquire an accurate 3D model in case we assume known geometry as input. We continue by explaining what is surface reflectance and scene illumination and how they are used in this thesis. We conclude by introducing the deep learning model that is used for geometry, reflectance or illumination estimation as well as the datasets required to produce training data for our CNN architectures.

2.1 Photometric Image Formation

As briefly explained in Chapter 1, an - RGB - image describes how interactions between the geometry, surface properties and environmental lighting in a scene generate through the camera optics and sensor properties a projection of 3D geometric features in the world into 2D features in the image made up of discrete color or intensity values. Fully modeling the image formation process would require a complex, multi-parameter model that takes into account all the effects that occur in each step of this process which is out of the scope of this thesis. Instead, we adopted a simplified model [142] of the photometric image formation: Light emitted by one or more light sources in the scene hits the

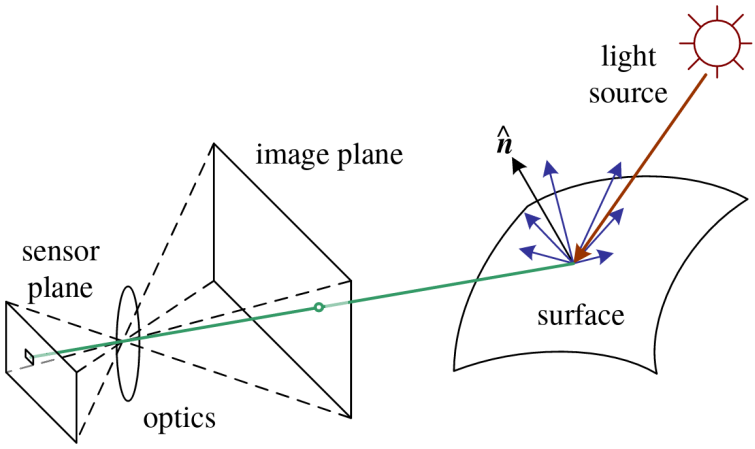


Figure 2.1: The adopted model [142] for photometric image formation: light emitted by one or more light sources in the scene hits the surface of one or more objects with specific geometry and material properties and passing through the camera's optics (lenses), it reaches the imaging sensor and is consequently converted into the digital (R, G, B) values. Image taken from [142].

surface of one or more objects with specific geometry and material properties and passing through the camera's optics (lenses), it finally reaches the imaging sensor and is consequently converted into the digital (R, G, B) values that we observe when we look at a digital image. Fig. 2.1 shows this process.

To this point we remind that the goal of this thesis, as defined in Sec. 1.3, is to undo the photometric image formation, *i.e.* decompose a single image or a sequence of images depicting an object in a scene into its intrinsic geometry, material and lighting. Without any assumptions or prior information, however, factorizing an image into these intrinsic components is a very difficult and under-constrained task, as the same visual result might be due to many different combinations of intrinsic components. For example, an image depicting a red sphere under white light might as well visualize a white sphere under red light. In the following paragraphs we introduce the assumptions made for each component participating in the image formation process. More details can be found in the next sections.

Geometry An object in the scene is represented by a 3D model

$$M = \{\mathbf{p}_i, \mathbf{n}_i\} \mid i = [1 \dots N], \tag{2.1}$$

that consists of N 3D points $\mathbf{p}_i$ with their corresponding normals $\mathbf{n}_i$. In this representation, every 3D point is a vector $\mathbf{p} \in \mathbb{R}^3$ that shows the point's position and every normal is a vector $\mathbf{n} \in \mathbb{R}^3$ that defines the orientation of the underlying surface in that position.

Material The material properties of a surface are efficiently represented using the BRDF, as originally defined in Sec. 1.2.2. As explained, this function describes how light is reflected at an opaque surface, essentially defining how diffuse, specular, etc a material is (*e.g.* a metallic surface is very specular and has a mirror-like reflection while a surface made of fabric is mostly diffuse). Typically, BRDFs have been approximated using parametric modeling (*e.g.* physical modeling, heuristic modeling, empirical observations). In this thesis, however, we go beyond these simplifying models and assume a non-parametric representation for the BRDF (*i.e.* data-driven BRDFs).

Lighting In the simplified photometric image formation model described above, the light finally reaching the camera's sensor comes from one or more light sources in the scene. As explained in Sec. 1.2.3, this accounts for direct lighting coming from different directions in the scene but is a simplification since it does not take into account any indirect lighting, meaning light bounces in the scene elements that turn the latter into light sources too. In this thesis, when referring to environmental (or natural) illumination we assume *environment maps*. An environment map is an omni-directional (360° panoramic) HDR image that accounts for natural lighting coming from every direction (*e.g.* a real scene with sky, windows, etc), without including multiple light bounces. The use of environment maps in rendering allows for high realism without prohibitively increasing the computation time.

Optics Once the light from a scene reaches the camera, it must still pass through the lens before reaching the sensor. For our applications in this thesis, it suffices to treat the lens as an ideal pinhole that simply projects all rays through a common center of projection. However, if we want to deal with issues such as focus, exposure, vignetting, and chromatic aberration, a more sophisticated image formation model needs to be considered.

Digital camera Light falling on an imaging sensor is usually picked up by an *active sensing area*, integrated for the duration of the exposure (usually expressed as the shutter speed in a fraction of a second, *e.g.* $\frac{1}{125}$, $\frac{1}{60}$, $\frac{1}{30}$), and then passed to a set of *sense amplifiers*. The two main kinds of sensor used in digital still and video cameras today are CCD and CMOS. The main factors affecting the performance of a digital image sensor are the shutter speed, sampling pitch, fill

factor, chip size, analog gain, sensor noise, and the resolution (and quality) of the analog-to-digital converter. No special assumptions regarding the type of digital camera or the imaging sensor are made.

2.2 Object Geometry

In this section, we will describe tools and techniques for obtaining information about the geometry of 3D scenes from 2D images. This task is challenging because the image formation process, introduced in the previous section, is not generally invertible: from its projected position in a camera image plane, a scene point can only be recovered up to a one-parameter ambiguity corresponding to its distance from the camera. Hence, additional information is needed to solve the reconstruction problem. For uncalibrated setups, a popular solution is to use the motion of the camera to find corresponding image points in multiple views, namely SfM. For calibrated setups, PS allows for the estimation of surface normals from shading variations observed in images taken under differently-oriented illumination. In what follows, we briefly describe the basic principles of these two techniques that are later used in Chapter 3.

2.2.1 Structure-from-Motion

Humans perceive a lot of information about the 3D structure in their environment by moving through it. When the observer moves and the objects around the observer stay still, information is obtained from images sensed over time. The computer vision equivalent is SfM and assumes the use of a digital camera for estimating 3D structures from 2D image sequences that may be coupled with local motion signals [101].

The first step in SfM is to find correspondences between the images and the reconstruction of the 3D object. To do so, features such as corner points (edges with gradients in multiple directions) are tracked from one image to the next. One of the most widely used feature detectors is the Scale-Invariant Feature Transform (SIFT) [95]. It uses the maxima from a Difference-of-Gaussians (DoG) pyramid as features. The first step in SIFT is finding a dominant gradient direction. To make it rotation-invariant, the descriptor is rotated to fit this orientation. Another common feature detector is the Speeded Up Robust Features (SURF) [11]. In SURF, the DoG is replaced with a Hessian matrix-based blob detector. Also, instead of evaluating the gradient histograms, SURF computes the sums of gradient components and of their absolute values.

Figure 2.2: ARC3D: A web service, including a group of SfM tools developed at KU Leuven, that allows non-expert users to upload digital images (right-top), perform a 3D reconstruction of the desired scene (right-bottom) and download the output 3D model (left).

In a second step, the features detected from all the images are matched. One of the matching algorithms that track features from one image to another is the Lukas-Kanade tracker [96]. Sometimes some of the matched features are incorrectly matched. This is why the matches should also be filtered. RANSAC (Random Sample Consensus) is the algorithm that is usually used to remove the outlier correspondences. In the paper of Fischler and Bolles [46], RANSAC is used to solve the Location Determination Problem (LDP), where the objective is to determine the points in space that project onto an image into a set of landmarks with known locations.

In a final step, the feature trajectories over time are used to reconstruct their 3D positions (3D point-cloud) and the camera's motion (intrinsic and extrinsic parameters) [117]. As an optional post-processing step, dense depth matching [149] can be used to move from the sparse point-cloud to a dense 3D mesh.

In this thesis, whenever SfM is employed we use the ARC3D web service [148]. An example can be seen in Fig. 2.2. ARC3D includes a group of SfM tools, developed at KU Leuven, that allows non-expert users to upload digital images, perform a 3D reconstruction of the desired scene (*e.g.* a statue) and download the output 3D model.

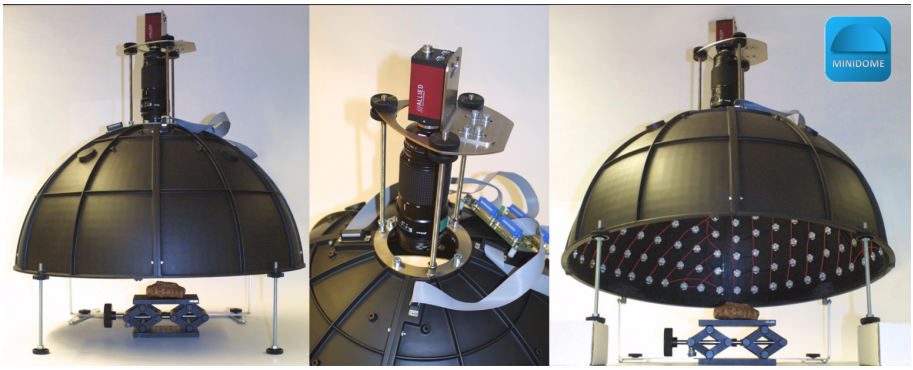


Figure 2.3: Minidome: An automated 3D digitizing solution, developed at KU Leuven, that uses principles from PS to capture geometric and radiometric information, while at the same time optimizing ease of use and flexibility.

2.2.2 Photometric Stereo

PS is a computer vision technique for estimating the surface normals of objects by observing that object under different lighting conditions. It is based on the fact that the amount of light reflected by a surface is dependent on the orientation of the surface in relation to the light source and the observer. By measuring the amount of light reflected into a camera, the space of possible surface orientations is limited. Given enough light sources from different angles, the surface orientation may be constrained to a single orientation or even overconstrained. Fig. 2.3 shows an example of a device (Minidome), developed at KU Leuven, that uses the PS technique to estimate surface normals.

The PS technique was originally introduced by Woodham [159]. Under Woodham's original assumptions - Lambertian reflectance, known point-like distant light sources, and uniform albedo - the problem can be solved by inverting the linear equation

$$\mathbf{I} = \mathbf{L} \cdot \mathbf{n}, \tag{2.2}$$

where $\mathbf{I}$ is a (known) vector of m observed intensities, $\mathbf{n}$ is the (unknown) surface normal, and $\mathbf{L}$ is a (known) $3 \times m$ matrix of normalized light directions. This model can easily be extended to surfaces with non-uniform albedo, while keeping the problem linear. Taking an albedo reflectivity of k, the formula for the reflected light intensity becomes

$$\mathbf{I} = k(\mathbf{L} \cdot \mathbf{n}). \tag{2.3}$$

If $\mathbf{L}$ is square (there are exactly 3 lights) and non-singular, it can be inverted, giving

$$\mathbf{L}^{-1}\mathbf{I} \;=\; k\mathbf{n}. \tag{2.4}$$

If $\mathbf{L}$ is not square (there are more than 3 lights), a generalization of the inverse can be obtained using the Moore-Penrose pseudoinverse by simply multiplying both sides with $\mathbf{L}^T$ giving

$$(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}\mathbf{I} \;=\; k\mathbf{n}. \tag{2.5}$$

After which the normal vector and albedo can be solved as described above.

The special case where the data come from a single view is known as *shape from shading*, and was analyzed by Horn [66]. The same principles can be applied to multiple views, however, resulting in Multiview PS [43], where starting from a rough initial shape the photometrically estimated normals can be used to refine the object's geometry.

The classical PS approaches concern themselves only with Lambertian surfaces, with perfectly diffuse reflection. This is unrealistic for many types of materials, especially metals, glass and smooth plastics, and will lead to aberrations in the resulting normal vectors. In this thesis, we go beyond these simplifying assumptions and we show how to refine geometry and estimate non-parametric reflectance for highly reflective materials using principles from PS. For more details regarding the latter the reader can visit Chapter 3.

2.3 Surface Reflectance

Comprehending the visual world around us requires understanding the role of *materials*. In essence, we think of the appearance of a material as being a function of how that material interacts with light. The material may reflect light or may exhibit more complex phenomena such as subsurface scattering. This 'lighting' behavior of the material is characterized as *reflectance*. In literature many different models have been developed for reflectance. Below we first describe its most general form, the BRDF, and then look at some more specialized models, like the diffuse, specular and Phong shading models.

2.3.1 Bidirectional Reflectance Distribution Function

When light hits an object's surface it is scattered and reflected (see Fig. 2.4a). The most general model that describes this interaction between the light and the surface is the *BRDF*. Relative to some local coordinate frame on the surface

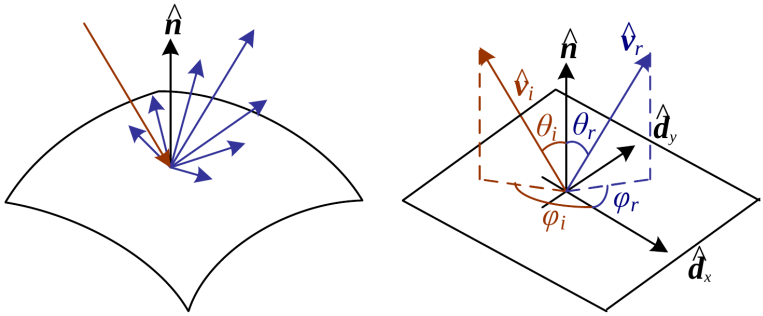


Figure 2.4: (a) When light hits a surface, it is scattered and reflected. (b) The BRDF $f(\theta_i, \phi_i, \theta_r, \phi_r)$ models this interaction between the light and the surface and is parameterized by the angles that the incident, $\mathbf{v}_i$, and reflected (or outgoing), $\mathbf{v}_r$, light ray directions make with the local surface coordinate frame $(\mathbf{d}_x, \mathbf{d}_y, \mathbf{n})$. Image courtesy of [142].

$(\mathbf{d}_x, \mathbf{d}_y, \mathbf{n})$, the BRDF is a 4D function that defines how much of each wavelength λ arriving at an *incident* direction $\mathbf{v}_i$ is emitted in a *reflected* (or *outgoing*) direction $\mathbf{v}_r$ (see Fig. 2.4b). The BRDF is usually written as a function of the polar angles of incident and reflected light relative to the surface frame, as:

$$f(\omega_i \rightarrow \omega_r) \;=\; f(\theta_i, \phi_i, \theta_r, \phi_r) \tag{2.6}$$

2.3.2 BRDF Properties

Before we look at specific BRDF models, let us discuss a few properties shared by all BRDFs. The first is *energy conservation*, and it refers to the fact that it is impossible for a surface to reflect more light than was incident on it, because all incident light must either be reflected or absorbed, and no light may be created during reflection. As a mathematical expression, this means that the integral of the BRDF over all outgoing directions, scaled by a cosine term to account for foreshortening, must be less than one,

$$\forall \omega_i : \int_{\Omega} f(\omega_i, \omega_r) cos\theta_r d\omega_r \leq 1. \tag{2.7}$$

A second more subtle property of the BRDFs, called *Helmholtz reciprocity*, is due to the symmetry of light transport and it refers to the fact that BRDFs must be unchanged when the angles of incidence and exitance are swapped,

$$f(\omega_i \rightarrow \omega_r) \;=\; f(\omega_r \rightarrow \omega_i). \tag{2.8}$$

The term *physically-plausible BRDF* is sometimes used for reflectance functions that satisfy energy conservation and reciprocity. Some, but not all, BRDFs have a property called isotropy, meaning that they are unchanged if the incoming and outgoing vectors are rotated by the same amount about the surface normal. This practically means that the BRDF is a 3D function in this case, depending only on the difference between the azimuthal angles of incidence and exitance,

$$f(\omega_i \rightarrow \omega_r) \;=\; f(\theta_i, \theta_r, |\phi_r - \phi_i|). \tag{2.9}$$

The inverse of isotropy is *anisotropy*. However, anisotropy as well as other more complicated BRDF properties (*e.g.* asperity scattering, retro-reflection) are not studied in this thesis.

2.3.3 BRDF Models

We now turn to looking at specific examples of BRDF models. We examine simple examples where the reflectance is expressed as a mathematical formula. Real surfaces, of course, are more complex and mathematic formulas frequently do not predict the reflectance with great accuracy.

Lambertian BRDFs The simplest possible BRDF is a constant

$$f \;=\; const. \tag{2.10}$$

This results in a matte or diffuse appearance, and is known as ideal Lambertian reflectance.

Phong BRDFs Another simple analytic BRDF is the Phong model [116], designed to qualitatively mimic the appearance of glossy materials,

$$f \;=\; k_s(\mathbf{r} \cdot \mathbf{v})^n, \tag{2.11}$$

where $\mathbf{v}$ is the view direction and $\mathbf{r}$ the mirror reflection of the light direction from the tangent plane.

In general, there are many parametric BRDF models in both computer vision and graphics, ranging from ad-hoc models (*e.g.* Blinn-Phong [15], Lafortune [78], Ashikhmin [6], DSBRDF [107]) designed for efficiency, to physics-based derivations either based on the micro-facet theory (*e.g.* Ward [156], Cook-Torrance [25], Schlick [130]) or wave optics (*e.g.* He [59]). For a comparison of various reflectance models we refer the reader to empirical studies like [105].

Non-parametric BRDFs Although we could continue to develop mathematical BRDF formulas of increasing sophistication that explains a great variety of optical phenomena, over the past decade it has became increasingly practical to simply measure the BRDFs of real material samples [99]. In fact, this is one main avenue of research surveyed in this thesis: that measured data can capture a greater variety of real-world optical phenomena with greater accuracy than is possible with analytic models. For more details we refer the reader to Chapter 4.

2.4 Scene Illumination

Images can not exist without light. In particular, to produce an image, the scene must be illuminated with one or more light sources. Light sources can generally be divided into point and area light sources.

2.4.1 Point Light Sources

A point light source originates at a single location in space (*e.g.* a small light bulb), potentially at infinity (*e.g.* the sun). In addition to its location, a point light source has an intensity and a color spectrum, *i.e.* a distribution over wavelengths $L(\lambda)$. The intensity of a light source falls off with the square of the distance between the source and the object being lit, because the same light is being spread over a larger (spherical) area. A light source may also have a directional falloff (dependence), but we ignore this here.

2.4.2 Area Light Sources

Area light sources are more complicated. A simple area light source, such as a fluorescent ceiling light fixture with a diffuser, can be modeled as a finite rectangular area emitting light equally in all directions.

2.4.3 Environment maps

The simple shading models presented thus far (*i.e.* point and area light sources) assume that light rays leave the light sources, bounce off surfaces visible to the camera with specific material properties, thereby changing in intensity or color, and arrive at the camera (*direct illumination*). In reality, light sources can be

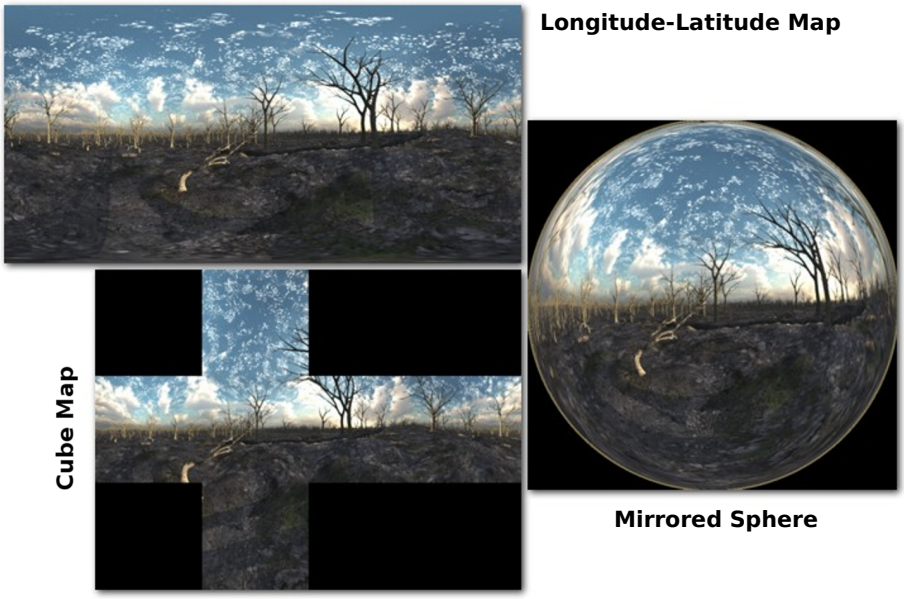


Figure 2.5: An environment map is captured by taking a HDR image of a mirrored sphere (right) and unwrap it onto a longitude-latitude map (left-top) or a cube map (left-bottom). This procedure is described in [29].

shadowed by occluders and rays can bounce multiple times around a scene while making their trip from a light source to the camera (*indirect illumination*).

In this thesis, although we do not explicitly handle the indirect illumination, we work with complex light distributions that approximate, say, the incident illumination on an object sitting in an outdoor scene, called *environment maps*. This representation maps incident light directions $\mathbf{v}_i$ to color values, $L(\mathbf{v}_i)$, and is equivalent to assuming that all light sources are at infinity. Environment maps can be represented as a collection of cubical faces [55], as a single longitude-latitude map [16], or as the image of a reflecting sphere [157]. A convenient way to get a rough model of a real-world environment map is to take an HDR image of a reflective mirrored sphere and to unwrap this image onto the desired environment map [29]. Fig. 2.5 shows such an example. For more details about environment mapping, including the formulas to map directions to pixels for the three most commonly used representations, we refer the reader to [157].

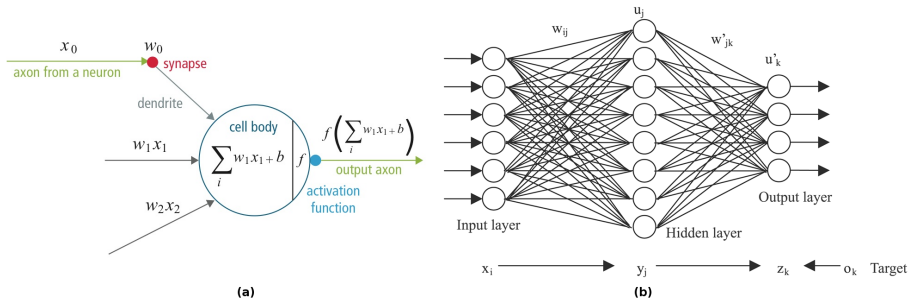


Figure 2.6: CNNs (a) consist of several basic components, the neurons, and (b) are usually organized in convolutional or fully connected layers, before reaching a loss layer. They are the dominant method for many computer vision tasks.

2.5 Convolutional Neural Networks

The last years a machine learning approach, called *deep learning*, has emerged and has dominated in many computer vision tasks, such as image classification and object detection. Deep learning is not a new technique. In fact, its origin traces back to the 1970s and 1980s, but due to the lack of GPU processing power it was impossible to be used at its full potential back then. Fast forward to 2010s, when Krizhevsky *et al.* [76] won the ImageNet classification competition [31] by a large margin and deep learning became a trend in computer vision and machine learning conferences. In this thesis, we focus on a specific type of deep learning, the *CNNs*, for estimating geometry, reflectance or illumination.

CNNs consist of multiple layers of *receptive fields*. These are small *neuron* collections which process portions of the input image. Each neuron receives a number of inputs and produces an output. In particular, the neuron calculates the weighted sum of its input values and then it applies an activation function (typically a sigmoid) to the output (Fig. 2.6a). Generally, CNNs can be seen as combinations of *convolutional* and *fully connected* layers, with pointwise nonlinearity applied at the end of or after each layer, that end up to a *loss* layer (Fig. 2.6b). In the following paragraphs, we describe the most common types of CNN layers as well as a famous architecture used in the thesis.

2.5.1 CNN Layer Types

A CNN architecture is formed by a stack of distinct layers that transform the input volume into an output volume (*e.g.* holding the class scores) through a

differentiable function. A few distinct types of layers are commonly used.

Convolutional layer The convolutional layer is the core building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field, but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input and producing a 2-dimensional activation map of that filter. As a result, the network learns filters that get activated when it detects some specific type of feature at some spatial position in the input. Stacking the activation maps for all filters along the depth dimension forms the full output volume of the convolution layer. Every entry in the output volume can thus also be interpreted as an output of a neuron that looks at a small region in the input and shares parameters with neurons in the same activation map.

Pooling layer Another important concept of CNNs is pooling, which is a form of non-linear down-sampling. There are several non-linear functions to implement pooling among which *max pooling* is the most common. It partitions the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. The intuition is that once a feature has been found, its exact location isn't as important as its rough location relative to other features. The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. It is common to periodically insert a pooling layer in-between successive convolutional layers in a CNN architecture. The pooling operation provides a form of translation invariance. In addition to max pooling, the pooling units can also perform other functions, such as *average pooling* and even *L2-norm pooling.*

ReLU layer ReLU is the abbreviation of Rectified Linear Unit. This is a layer of neurons that applies the non-saturating activation function $f(x) = max(0, x)$. It increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer. Other functions are also used to increase nonlinearity, *e.g.* the hyperbolic tangent $f(x) = tanh(x)$ or the sigmoid function $f(x) = (1 + e^{-x})^{-1}$, but compared to these functions the usage of ReLU is preferable, because it results in the neural network training several times faster, without making a significant difference to generalization accuracy.

Fully connected layer Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected

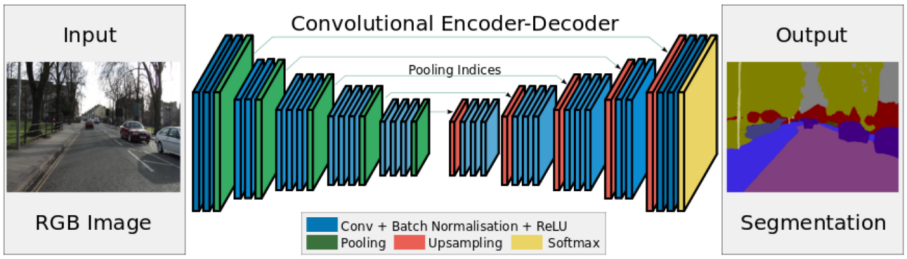


Figure 2.7: A typical architecture of the convolutional encoder-decoder network. This image originates from [7].

layers. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in Fig. 2.6b. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

Loss layer The loss layer specifies how the network training penalizes the deviation between the predicted and true labels and is normally the last layer in the network. Various loss functions appropriate for different tasks may be used there. *Softmax loss* is used for predicting a single class of K mutually exclusive classes. *Sigmoid cross-entropy loss* is used for predicting K independent probability values in $[0, 1]$. *Euclidean loss* is used for regressing to real-valued labels $[-\infty, \infty]$.

2.5.2 Convolutional Encoder Decoder

In Chapters 5 and 6, we discuss the use of CNNs for estimating per-pixel normals, BRDF parameters and natural illumination. The proposed architectures are based on the *convolutional encoder-decoder* architecture [7]. The latter consists of a sequence of non-linear processing layers (encoders) and a corresponding set of decoders followed by a pixelwise classifier. Typically, each encoder consists of one or more convolutional layers with batch normalization and a ReLU non-linearity, followed by non-overlapping max pooling and sub-sampling. The sparse encoding due to the pooling process is upsampled in the decoder using the max pooling indices in the encoding sequence (see Fig. 2.7 for an example). One key ingredient of the convolution encoder-decoder is the use of max-pooling indices in the decoders to perform upsampling of low resolution feature maps. This has the important advantages of retaining high frequency details in the segmented images and also reducing the total number of trainable parameters

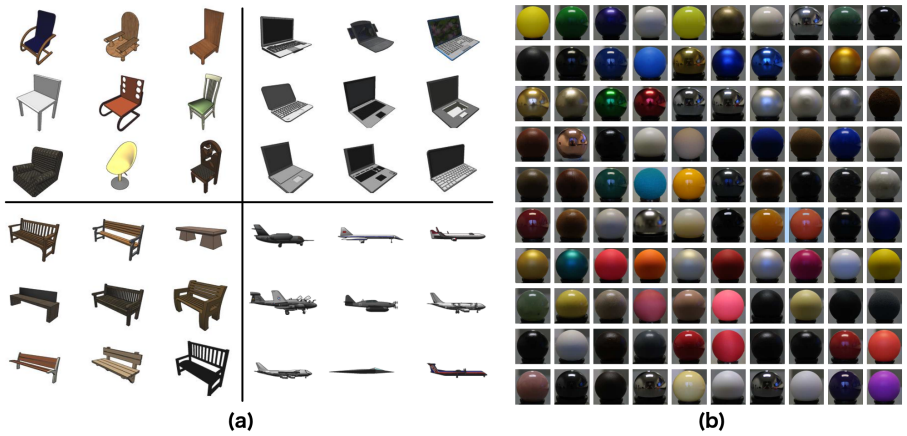


Figure 2.8: (a) Example 3D shapes from the ShapeNet dataset [22]. (b) Example BRDF samples from the MERL BRDF database [99]. These images originally come from the corresponding publications.

in the decoders. The entire architecture is usually trained end-to-end using stochastic gradient descent.

2.6 Datasets

In order to train the CNN architectures introduced in the previous section a large number of training data is required (in the order of tens or hundreds of thousands of images). Usually this data comes from images that researchers collect by themselves or even from the web. However, when the acquisition procedure is very difficult or time consuming, a popular alternative has been the generation of synthetic data. To do so, we rely on repositories that contain scanned 3D objects, surface materials and environment maps. In what follows, we describe the datasets used in this thesis to generate synthetic images for training our CNN architectures presented in Chapters 5 and 6.

2.6.1 ShapeNet

ShapeNet [22] is a richly-annotated, large-scale repository of shapes represented by 3D CAD models of objects. ShapeNet contains 3D models from a multitude of semantic categories and organizes them under the WordNet taxonomy. Fig. 2.8a

shows example 3D shapes from four semantic categories, chairs, laptops, benches, and airplanes. It is a collection of datasets providing many semantic annotations for each 3D model such as consistent rigid alignments, parts and bilateral symmetry planes, physical sizes, keywords. Annotations are made available through a public web-based interface. In this thesis, ShapeNet is our source of 3D car models used to generate synthetic training data in Chapters 5 and 6.

2.6.2 MERL BRDF Database

The *MERL BRDF database* [99] contains reflectance functions of 100 different materials measured using a gonioreflectometer. In particular, a light source illuminates the material to be measured from a hemisphere around it and a sensor then captures the light reflected from that material. The hemispherical rotation dimensions of the sensor and light source are the four dimensions of the BRDF, as described in Sec. 2.3.1. Each reflectance function is consequently stored as a densely non-parametric BRDF. Fig. 2.8b shows example materials from the database. The MERL BRDF database is our source of material BRDFs in the four following chapters of this thesis.

2.7 Conclusion

This chapter presented the background knowledge that the following chapters built upon. Since the topic of this thesis is about the decomposition of one or more images into their intrinsic geometry, material and lighting we start by describing the image formation process in the camera. We then proceed to methods for extracting 3D shape from 2D images and in particular we present the SfM and PS techniques that are combined in Chapter 3. Next, we discuss the basic principles of surface reflectance and scene illumination that are used in Chapters 4 and 5 respectively. Finally, we introduce the CNN models and datasets required for estimating per-pixel normals, Phong BRDF parameters and environment maps in Chapters 5 and 6.

Chapter 3

Extracting 3D Shape and Surface Reflectance

PS and SfM are two well researched methods for image-based 3D reconstruction, as explained in Sec. 2.2. In general, PS excels in reconstructing high-frequency surface details, while SfM is superior for reconstructing the overall, low-frequency 3D shape. Both of these methods are better tailored towards non-specular surfaces. In literature, recent works have combined the two approaches, in order to get the best of both worlds. Typically, they assume simple reflectance models, like Lambertian behavior possibly mixed with some specular lobes (see Sec. 2.3), and they focus on shape rather than reflectance. Most importantly, the data capture is performed with sophisticated hardware setups that are expensive and inaccessible to most researchers, let alone casual users.

In this chapter, we investigate the use of simple flash-based photography to deal with the problem of capturing an object's 3D shape and surface reflectance characteristics at the same time, which is the core of this thesis. In particular, we present our approach to tackle the Research Question 1 introduced in Sec. 1.3: *Can we extract 3D shape and surface reflectance from a small set of uncalibrated images and under which lighting conditions?* The presented method combines the principles of SfM, MvS and PS, yet, we make sure not to use more than readily available consumer equipment, like a camera with flash or a smartphone. Starting from a low-resolution mesh generated from SfM and MvS, we apply a PS-based technique to refine both geometry and reflectance, where the latter is expressed in terms of data-driven BRDF representations.

The work presented in this chapter is based on[1]:

- S. Georgoulis, M. Proesmans and L. Van Gool, *Tackling Shapes and BRDFs Head-on.* Published in IEEE International Conference on 3D Vision (3DV) 2014.

3.1 Introduction

The real world is flooded by objects made of different materials and placed at different natural environments. The appearance of such an object is essentially determined by three independent factors: (1) its geometry (*i.e.* 3D shape), (2) its surface materials (*i.e.* surface reflectance) and (3) the lighting environment that the object is currently observed in (*i.e.* environment map). In order to reproduce this appearance at the same or a different lighting environment, one has to decouple it into its geometric (3D shape) and photometric (surface reflectance) properties. Though challenging, this procedure is important in the context of various application domains such as advertisement, movie production and cultural heritage preservation where photo-realistic images of real objects need to be synthesized. As a result, both a high-quality 3D shape of the object as well as a detailed model of reflectance across its surface need to be acquired. Yet, it is fair to say that in the last decades 3D shape extraction has progressed far more than the extraction of reflectance.

Indeed, when it comes to extracting 3D shapes, steady progress in the quality obtained with uncalibrated SfM (i.e. 3D modeling on the basis of images taken with a hand-held camera with unknown settings [145]) has brought this technology close to if not at par with what laser scanners, another impressive technology, can achieve. Furthermore, several PS and MvS -based methods [169, 111, 110] have been proposed that take advantage of material characteristics to improve results where traditional SfM might fail.

In contrast, systems that provide a reflectance model that comes somewhat close to the ideal BDRF [106] at the different surface points are by and large lacking still. Such BRDF has to be recorded with dedicated devices called gonioreflectometers that independently drive a light source and a sensor to many different positions around the object [98, 99, 64]. Nevertheless, a dense sampling of an object's BRDF - usually only of a small planar patch - is a time-consuming process. As estimated in [74], for a sampling at an angular

[1]This work has been extended and included in the following paper: S. Georgoulis, V. Vanweddingen, M. Proesmans and L. Van Gool, *Shape and Reflectance Using a Camera with Flash.* Submitted in IEEE International Journal on Computer Vision (IJCV).

resolution of 1 degree more than 10^8 measurements are required. Instead, most systems go for a far simpler reflectance model, like Lambertian behavior, possibly mixed with some specular lobes.

Therefore, joint 3D shape and surface reflectance capture remains an important and challenging problem. On the one hand, this capture is performed with dedicated hardware setups situated in laboratory environments such as the light stage of Ghosh *et al.* [52] and the coaxial lights of Holroyd *et al.* [64]. Although these methods achieve accurate results, the data capture is performed with sophisticated hardware setups that have proven to be complicated, expensive and most importantly inaccessible to casual users. On the other hand, recent approaches [110] moved outside the darkroom in order to estimate 3D shape and surface reflectance "in the wild". Even in such cases though, one has to record HDR images (meaning multiple exposures per viewpoint), also scan the environment lighting (per viewpoint) and finally carefully align the two (for each viewpoint) using non-trivial calibrating techniques. As a result, the scanning procedure becomes impractical for casual users.

In this chapter, we propose a simple capturing setup and a joint reflectance and geometry refinement technique to remedy this situation. Compared to existing approaches our system fulfills three basic principles in order to bring shape and reflectance acquisition closer to non-expert users: (1) only readily available consumer equipment is required for the scanning, like a flash-equipped DSLR camera or a smartphone, (2) only a small number of LDR images need to be recorded, (3) minimum to no intervention is required by the user (e.g. using calibration techniques, creating masks, etc). This is the first system providing to casual users the ability to extract both 3D shape and surface reflectance under such conditions.

Specifically, the user should simply walk around the object, taking photos under the illumination of the camera's flash, which is considered dominant over any other illumination in the scene. Using uncalibrated SfM and consequently MvS an initial 3D shape is extracted. The presented method then estimates photo-realistic lower-dimensional BRDFs (BRDF slices) from sampled sections[2], and uses the estimates to refine both reflectance and geometry based on the proposed data-driven optimization technique. Fig. 1.1 gives a preview.

After reviewing related work in Sec. 3.2, a system overview is presented in Sec. 3.3. The input assumptions and initial geometry are discussed in Sec. 3.4. Our data-driven reflectance model is introduced in Sec. 3.5, where we explain how base material BRDFs are extracted from clustered regions of similar reflectance. In Sec. 3.6 we explain how base material BRDFs, photometric normals, material

[2]Since the viewpoint/light distance is fixed, the BRDF is only sparsely sampled and these samples are concentrated in a slice of the BRDF domain (see Sec. 3.4)

weights and 3D points are refined. Finally, we analyze the results for a number of challenging objects, both synthetic and real, in Sec. 3.7 and conclude the chapter in Sec. 3.8.

3.2 Previous Work

The amount of research performed on 3D shape and surface reflectance acquisition is vast. In what follows, we organize an overview by topic:

3D shape acquisition PS [159] computes surface normals from shading variations observed in images taken under differently-oriented illumination. The classical formulation, as described in Sec. 2.2.2, assumes perfect Lambertian reflectance, but recent efforts have been directed towards robustification against outliers, such as shadows and specular highlights [19, 151, 111]. The key insight here is that most specular materials exhibit an approximately Lambertian behaviour for at least a subset of the viewing/lighting combinations. By singling out this matte component, Lambertian PS is still applicable in many cases. For setups like ours though, with only one lighting direction per viewpoint, finding this matte component becomes challenging, especially as the number of input images degrades. Other methods [60, 54, 61] are dedicated to capturing non-Lambertian phenomena such as specular highlights and iridescence. Our method also leverages the cues hidden in specular highlights, but in contrast to these works it is able to capture both 3D shape and surface reflectance from the same set of images.

Traditional PS [159] produces a two-dimensional normal field that may be integrated into a depth map [26, 19]. Using MvS, traditional PS can be extended to multiple views [43, 160, 111] in order to recover a full 3D representation of the scene. Such methods often rely on a base surface obtained from MvS which is then refined using shading information. However, all these methods restrict themselves to Lambertian objects of constant albedo. Due to the use of data-driven BRDF representations our approach generalizes better to a broader range of surfaces.

Image-based modeling methods, apart from a 3D shape, can also reconstruct a 'texture map' to model objects. In most recent methods, like [85, 49], texture color at each point is decided according to its image back-projections. Even then, a fixed texture map is insufficient to represent reflectance properly (we refer the reader to Fig. 1.1 for an indicative example).

Flash-aided reconstruction Flash-aided reconstruction offers a number of advantages, like facilitating data capture under low ambient light, having minimum number of self-shadows and reasonably-controlled illumination. Several works [115, 39, 2] fuse flash/no-flash image pairs to remove undesired artifacts like noise and specular highlights. Yet, if one wants to keep image capturing simple and use a hand-held camera as proposed in this chapter, the need for such image pairs taken from the same viewpoint entails a hard registration issue.

Melendez *et al.* [100] use images taken with a flash to provide shading information for Lambertian PS. Rather than avoiding the specular highlights produced by the flash, Lanman *et al.* [80] exploit them via a calibrated multi-flash system to obtain shape and reflectance using SfM and PS. In general, our data-driven approach makes far weaker assumptions both on geometry and reflectance and does not require any calibration target in the scene (something non-trivial for casual users).

Reflectance measurement A BRDF [106] describes the fraction of reflected light for all pairs of incoming/outgoing light directions. High-resolution BRDFs are traditionally acquired with a camera or a spectrometer, using dedicated hardware setups [27, 99, 45] that automatically capture many measurements covering the whole BRDF domain. Yet, they assume the 3D shape is known in advance. Methods like [33, 124] are related to ours but are only applicable to near-planar surfaces.

When a BRDF measuring device is not available, a fruitful alternative for reflectance acquisition "in the wild" is the base material representation introduced by Lensch *et al.* [83]. Since most objects consist of a limited number of materials, it is not necessary to separately model the full BRDF at each surface element individually. Instead, the reflectance of an element can be described as a combination of one or several base material BRDFs. Surface regions corresponding to the same material can be obtained by including reference objects in the images [60, 124] or by clustering the input data [83, 54, 147]. Since all surface points belonging to each base material region contribute to the same BRDF, the density at which those base material BRDFs can be sampled is far better than the density achievable for individual surface elements.

3D shape scanning and reflectance estimation Laser scanners and Structured Light (SL) patterns can be used to obtain accurate 3D shape [84, 127, 165]. Based on a precise 3D reconstruction, parametric reflectance models can then be fitted to each surface point based on image observations, as in [83, 91]. But solving the difficult problem of precisely registering images with

3D shape is crucial otherwise artifacts appear in misaligned regions. Combining reflectance from PS and shape from SL [165, 104, 3] solves this problem but the complicated setups coupled with non-trivial calibration procedures limit the method's applicability to casual users.

Joint 3D shape and reflectance estimation Because of the complementarity between shape and reflectance, quite some research effort has been put into jointly estimating shape and reflectance from imagery, or take advantage of the reflectance characteristics to improve the SfM result, using principles from PS. Methods like [54, 110] fit specific parametric BRDF models to input data, which may result in performance degradation when objects have a reflectance different from the assumed model.

Some other methods achieve accurate results by employing sophisticated hardware. Ma *et al.* [97] and Ghosh *et al.* [52] used a light stage with precisely controlled LED intensity, whereas Holroyd *et al.* [64] required coaxial lights. Both setups need expensive and complicated hardware to work. Some recent algorithms exploit various reflectance characteristics, like symmetries [4, 63] or monotonicity [135], to estimate shape and reflectance, but they either require up to a thousand input images [63] or rely on fragile optimization [3, 135]. Tan *et al.* [143] and Chandraker *et al.* [20] recovered iso-contours of depth and gradient magnitude for isotropic surfaces having the disadvantage that they need additional user interactions or boundary conditions to recover the shape.

Simpler setups Some attempts have been made to bring 3D shape and surface reflectance acquisition closer to casual users by employing simpler setups or by relaxing the lighting requirements. First, Higo *et al.* [62] used a custom-built hand-held camera with a single light source (Fig. 3.1a) to recover shape and reflectance from images but their approach relies on the Lambertian part of the material alone, and focuses on shape rather than material. Second, using a fixed mobile phone and a hand-held moving light tube (Fig. 3.1b), Ren *et al.* [124] acquired normal maps and reflectance but their approach is only usable for near-planar surfaces and it requires a BRDF chart to work. The latter practically means that the scanned materials are limited to the ones already existing in the BRDF chart. Third, Zhou *et al.* [169] have proven that the iso-contour approach for PS can be properly integrated with SfM, using a simple, yet fully-calibrated, lighting setup (Fig. 3.1c). Finally, most recently Oxholm and Nishino [110] have tried to relax the lighting requirements by moving outside the darkroom. Although, useful for shape and reflectance estimation "in the wild" their approach requires the capture of HDR images (i.e. multiple exposures have to be recorded for each viewpoint), the creation of masks for the shape initialization, the scanning of the environment map for each viewpoint

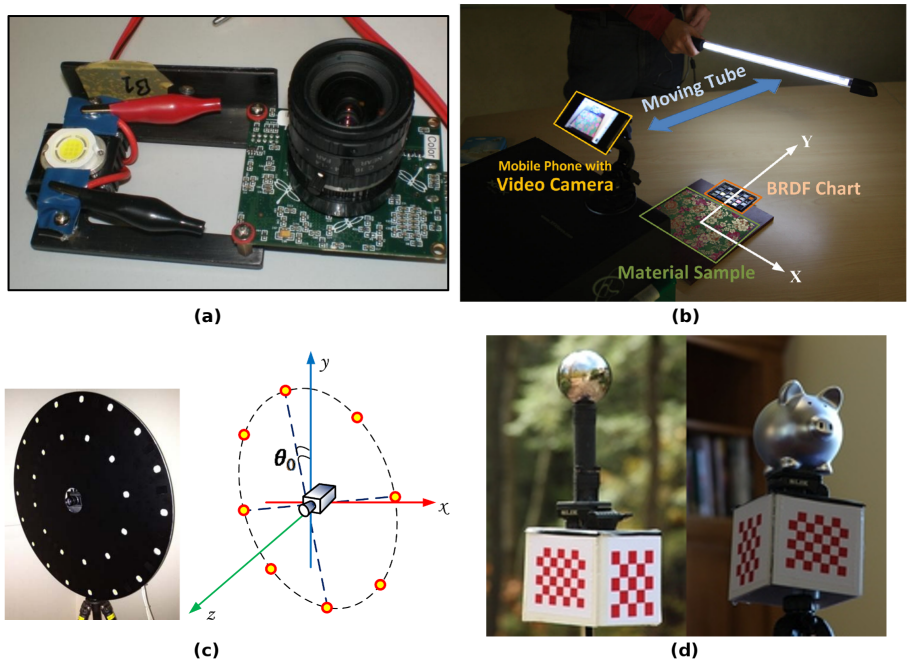


Figure 3.1: Attempts to extract 3D shape and surface reflectance using simple capturing setups. (a) Higo *et al.* [62] used a custom-build hand-held camera with a single light source. (b) Ren *et al.* [124] used a fixed mobile phone and a hand-held moving light tube. (c) Zhou *et al.* [169] used a simple, yet fully-calibrated, lighting setup. (d) Oxholm and Nishino [110] used a HDR DSLR camera together with a custom-made tripod.

and of course its alignment with the HDR image using the custom-made tripod of Fig. 3.1d. These pre-processing steps make the scanning rather impractical for non-expert users.

In contrast, our system is better tailored towards casual users for a number of reasons. It requires only readily-available consumer equipment, like a flash-equipped DSLR camera or smartphone, compared to custom-built setups that are not commercially available. Only a small number of LDR images need to be recorded. Going for too many images or HDR recordings would make the scanning process rather tiring. Lastly, minimum to no intervention is needed by the user. Tasks, such as lighting calibration, image registration or alignment, mask creation, etc, can be quite challenging for non-expert individuals.

3.3 System Overview

We investigate a simple setup consisting of a camera with built-in flash. The flash point of view is typically very close to the camera point of view. Compared to traditional PS approaches [169, 124, 43, 111] however, we will not have any reflectance information from light sources oblique to the camera point of view. As such, finding a subset of the viewing/lighting combinations where the scanned object exhibits an approximately Lambertian behaviour, as is the case with existing approaches, becomes challenging. Also, having both light and camera viewpoint almost aligned limits to undo the scale ambiguity in SfM [62]. If only one single light source is available, other assumptions are needed to recover reflectance information. We adopt the "base materials" approach that states that many objects are built out of a limited set of materials [83], meaning that one can still observe a material's reflectance changes by observing all points of the same material in the different camera viewpoints.

Given an image sequence, the system is initialized using uncalibrated SfM to recover the extrinsic and intrinsic camera parameters and reconstruct an initial 3D point cloud or mesh (see Sec. 3.4). While many methods combining SfM and PS, use isotropy and symmetry assumptions on the observed BRDFs [20, 4], or limit the computation to the Lambertian part of the observed reflectance [169, 62], in this chapter we investigate a data-driven approach, going straight for a proper BRDF estimation without any parametric or other assumptions, by observing the reflectance in the camera viewpoints, including specular highlights. The latter have proven to show intricate surface deviations that can not be observed by angles obeying the Lambertian law (*e.g.* [104, 43] vs [24, 167]). Here we want to remind that the overall goal is not only to arrive at an improved geometry and normal distribution, but also a plausible BRDF representation that can be used to create photo-realistic renderings of the resulting 3D model.

Every point in the mesh is characterized by its 3D position, its normal, and a set of weights related to a limited number of base materials, each with a BRDF representation. In this chapter we introduce a method where each of these parameters is to be refined by minimizing the color difference between the original and rendered images, using the BRDF estimations. Obviously, for our flash-based setup, we can not recover the full BRDF space, but we can recover a lower-dimensional BRDF (as will be discussed in Sec. 3.5). This is not necessarily a limitation though, since another line of our work presented later in Chapter 4 allows for the inference of the missing 2D/3D BRDF information from the 1D slice, enabling us to render the final model more confidently when the lighting conditions are different from the camera viewpoint.

3.4 Inputs, Assumptions and Initial Geometry

Inputs The input to our method is a sequence of M LDR images. Using SfM, both a sparse point-cloud as well as the intrinsic and extrinsic camera parameters are recovered. Consequently, MvS is used to arrive at a coarse mesh, which serves as an initialization for our method. However, methods based on turn-tables [47] can also be used for recovering the camera parameters and an initial mesh. In our experience though, the combination of SfM and MvS provides more accurate initializations over a visual-hull based approach [43], especially for objects with large concavities or complex topologies. Note that for the following experiments in Sec. 3.7, alternative approaches have been verified: (1) SfM + MvS, (2) Visual-hull + Poisson Surface Reconstruction (PSR) [72], (3) SfM + PSR. The latter cases show lower quality, but serve the purpose of verifying the sensitivity of the optimization process with respect to the initial mesh quality. Undeniably, the first approach (SfM + MVS) is generally more suitable for better results.

Assumptions There are no stringent requirements on the types of materials that can be processed, but generally for the SfM approach to work, they need to show a diffuse component and/or a reasonable amount of texture variation. Nevertheless, in Sec. 3.7 we tackle quite challenging examples showing a lot of reflectivity as well as texture-less smooth parts. Pure mirror surfaces have to be ruled out. It is useful to note that the flash is treated as a point light source and that we assume it to be dominant over other illumination in the scene; possible effects of environment lighting (e.g. [110]) have not been considered. For all real-life experiments we created linear images starting from the RAW images that were captured with the DSLR cameras that we used. Finally, for the smartphone setup the captured images were linearized too.

Initial geometry Our particular implementation of the SfM scheme [148], involves the detection of SURF-based features in each image, which are matched throughout the sequence. The feature correspondences determine the intrinsic parameters, pose estimations and 3D positions of these features. In a subsequent step, image pairs showing enough correspondences are rectified and disparity search is performed based on dynamic programming to create dense depth-maps. Using belief propagation on the depths observed in the original images, an initial dense mesh can be constructed. A preview of the whole pipeline, including the SfM step is shown in Fig. 3.2. Our SfM pipeline is publicly available as a web service[3], making it easily accessible to casual users. For the practical implementation of the consecutive multi-parameter optimization, the initial

[3]Link: http://www.arc3d.be/

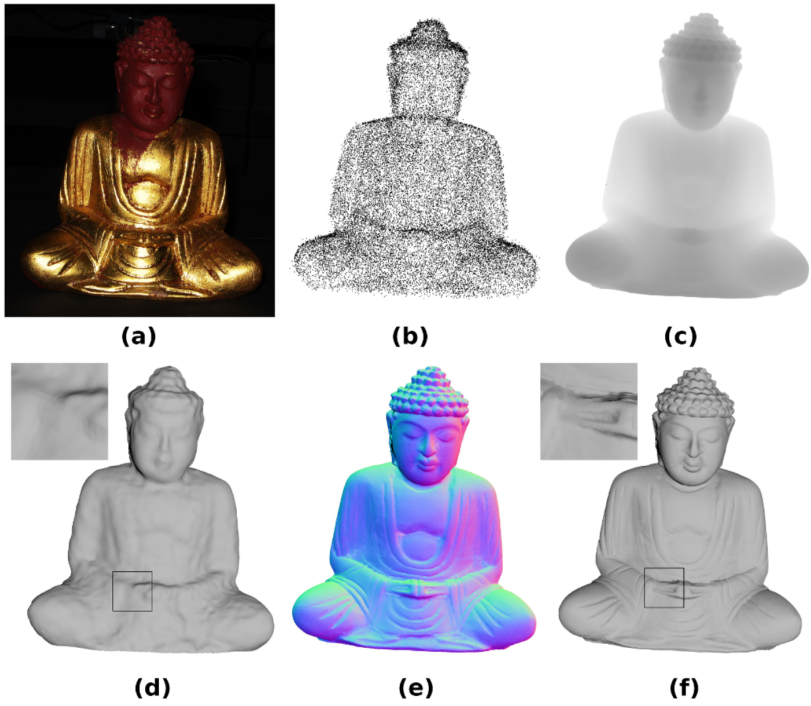


Figure 3.2: A preview of the proposed pipeline. (a) The user takes a sequence of pictures under the illumination of the camera's flash. (b) Using uncalibrated SfM, SURF-based 3D features are detected and matched throughout the sequence to determine the intrinsic and extrinsic camera parameters and 3D features' positions. (c) Next, image pairs are rectified to create dense depth-maps. (d) Using belief propagation on the depth maps an initial dense mesh is constructed. (e) The proposed method then samples and refines the base material BRDFs and uses them to estimate photometrically corrected normals. (f) Finally, the new normal estimates are used to refine the geometry.

mesh is re-sampled (uniform triangulation re-sampling) and further subdivided (if needed) to match the resolution of the images as seen from the camera viewpoints. Thus, the vertices of the initial mesh serve as a point-cloud for which both the base material BRDFs as well as the normals, 3D point positions and material weights are to be refined.

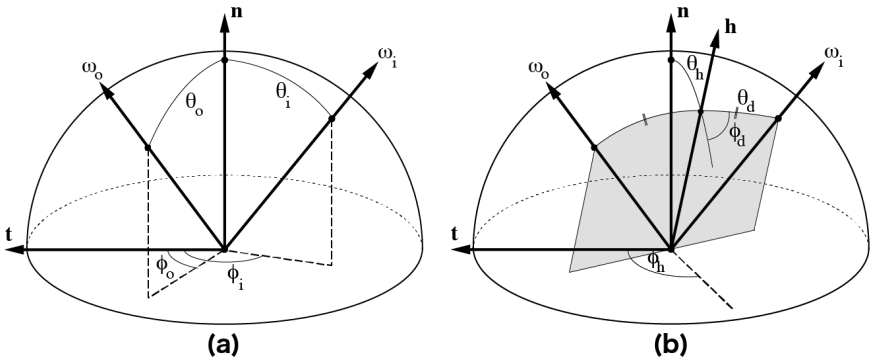


Figure 3.3: (a) The BRDF is a 4D function $f(\theta_i, \phi_i, \theta_o, \phi_o)$ of the spherical coordinates of the incoming light ω_i and the reflected light ω_o. (b) Using the re-parameterization of [128] that is based on the half vector $\mathbf{h}$, the BRDF is simplified to a 3D function $f(\theta_h, \theta_d, \phi_d)$.

3.5 Reflectance Model and Base Materials

3.5.1 BRDF Dimensionality

Before we enter into the discussion of the selected reflectance model, it is useful to introduce the concept of BRDFs in a bit more detail. Consider Fig. 3.3. It shows the lighting direction ω_i and the direction of observation (or reflection) ω_o. In Sec. 2.3.1 we show that specifying these directions fully, in order to express the percentage of directed incoming light that gets reflected into the direction of observation would take 4 angles in spherical coordinates (Fig. 3.3a). Thus, the corresponding BRDF would be a 4D function $f(\theta_i, \phi_i, \theta_o, \phi_o)$. Typically, people have used symmetry assumptions to simplify such a BRDF. For instance, one could consider the Rusinkiewicz re-parameterization [128] that is based on: (1) the half angle θ_h between the local surface normal $\mathbf{n}$ and the half vector $\mathbf{h}$ of the directions of light incidence ω_i and observation ω_o, and (2) the difference angle θ_d between the directions of incidence ω_i or observation ω_o and the half vector $\mathbf{h}$ (Fig. 3.3b). Most papers then use this pair of half angle θ_h and difference angle θ_d. For a broad range of isotropic surface materials these two angles suffice to generate a simplified 2D BRDF $f(\theta_h, \theta_d)$. Sometimes an anisotropic 3D BRDF $f(\theta_h, \theta_d, \phi_d)$ is used, by adding to (θ_h, θ_d) the angle ϕ_d, that specifies the rotation of the plane determined by ω_i and ω_o around the half vector $\mathbf{h}$. In principle, one should consider the BRDF per wavelength, which would add yet another dimension. In this thesis, we will work with three spectral bands as in cameras (red, green, blue) and extract a BRDF for each of those. As a matter of

fact, the BRDF can also be considered higher-dimensional if additional effects are taken into account, like spatial variations across a surface or for translucent objects where the place of light entry and exit can differ. These latter cases have not be considered in this thesis though.

3.5.2 Lower Dimensional BRDFs

Now consider a surface consisting of one single material. Assuming that a large range of normal directions can be seen from one single viewpoint and that the light is only coming from the flash, many surface points are viewed and lit under different angles. One single image can thus provide several samples of the material's BRDF. However, this will still not be enough to characterize the full anisotropic 3D BRDF of the material, as described above. Therefore, in our case we will consider a lower-dimensional BRDF representation, which will be used for both material clustering as well as BRDF re-sampling.

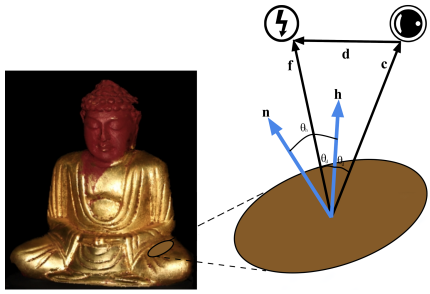


Figure 3.4: Local geometric configuration. As $\theta_d = const$, this leaves θ_h as the sole parameter of the BRDF.

Our starting point is the simplified 2D BRDF representation of [128], $f(\theta_h, \theta_d)$, that can effectively capture the behavior of the majority of isotropic materials. Fig. 3.4 shows our setup, in which $\mathbf{f}$ and $\mathbf{c}$ are the vectors of incoming (*i.e.* flash) and outgoing (*i.e.* camera) light respectively, and $\mathbf{n}$ is the surface normal. As explained above, the half vector $\mathbf{h}$ is the bisector of $\mathbf{f}$ and $\mathbf{c}$. In this way, θ_h measures the angular deviation from the direction of ideal specular reflection. Specular highlights appear around $\theta_h \approx 0$. This can be seen in the 2D BRDF visualization of Fig. 3.5, where the top row corresponds to $\theta_h = 0$. The abscissa represents θ_d.

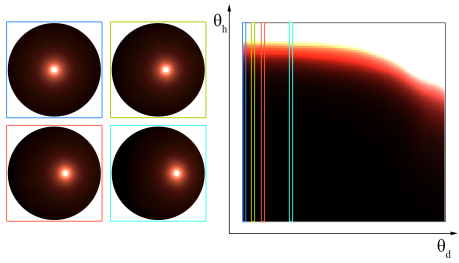


Figure 3.5: Lower dimensional BRDFs. For our fixed camera/flash setup, $\theta_d = const$. Our samples populate one of the first columns of the 2D BRDF.

With $\mathbf{B}$ as our per color-channel 2D BRDF, we model the reflected RGB color vector as:

$$\tilde{c} = \max(0, \mathbf{n} \cdot \mathbf{h})\, \mathbf{B}(\theta_h, \theta_d) \tag{3.1}$$

where the clipped dot product accounts for an attenuation due to an off-normal incident light direction (Lambert's cosine law). For our fixed camera/flash setup there is also the observation that ω_i and ω_o are very close. This practically means that not only θ_d is constant but also close to zero. The sole remaining parameter θ_h indicates the direction of the half vector relative to the local surface normal. In Fig. 3.5 our samples populate one of the first columns of the 2D BRDF, depending on the distance between the camera and the flash light $|\mathbf{f} - \mathbf{c}|$. Notice that this is not just *a* column of the 2D BRDF, but an important one since it contains vital information about the dynamic range. Indeed, for most materials the maximal intensity is obtained when the viewing, lighting, and surface normal directions are almost aligned. This column will be referred as 1D BRDF or BRDF slice.

3.5.3 Base Materials

To represent the BRDF, analytic models have been popular in literature, ranging from the Blinn-Phong [15] till the Directional Statistics BRDF model [107], but they usually involve the non-trivial guesstimation of initial parameters, while being susceptible to noise. Instead, our starting point is a purely data-driven approximation of a BRDF slice $\mathbf{B}(\theta_h)$, expressed as a θ_h-indexed vector of RGB values. Yet, we have to make sure not to mix up samples observed at points with different material properties. Therefore, we follow an approach similar to the one introduced by Lensch *et al.* [83]. We assume our surface to be composed of a finite number of base materials and let all surface points belonging to the same material contribute to that base material's BRDF.

3.5.4 Clustering into Base Materials

Consider an initial 3D model L consisting of N 3D points $\mathbf{P}$ and normals $\mathbf{n}$,

$$L \;=\; \{\mathbf{P}_i, \mathbf{n}_i\} \mid i = [1 \ldots N]. \tag{3.2}$$

We project every $\mathbf{P}_i$ into each of the camera viewpoints recovered from SfM, to obtain M pairs of RGB measurements ($\mathbf{c}$) and half angles (θ_h):

$$\mathbf{C}_i = \{\mathbf{c}, \theta_h\}_{ij} \mid j = [1 \ldots M],\ \zeta(i,j) = 1 \tag{3.3}$$

The term ζ_{ij} represents the visibility equating to one if $\mathbf{P}_i$ is visible in image j, and to zero otherwise. From each $\mathbf{C}_i$, we compute a low quality 1D BRDF, which

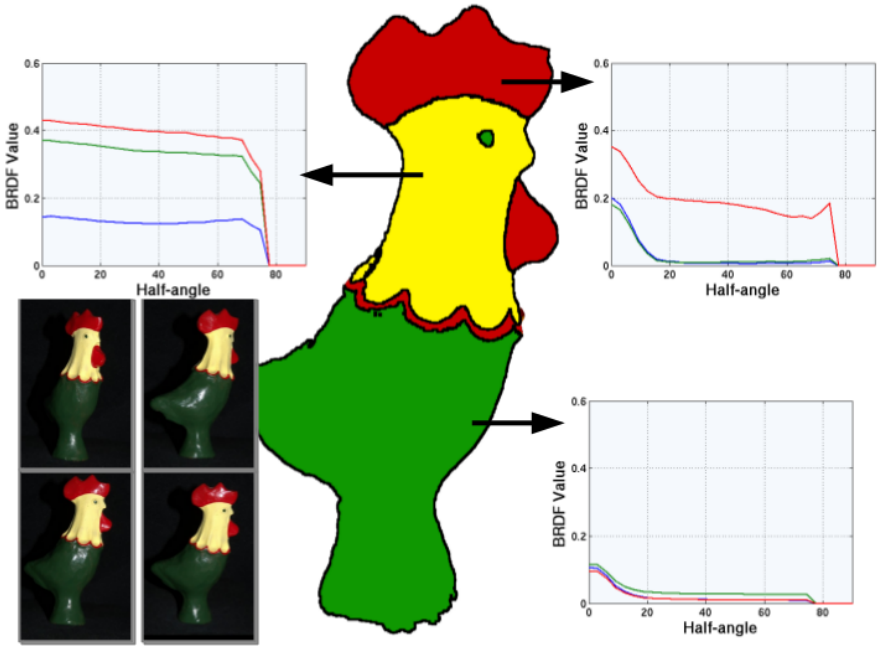


Figure 3.6: Clustering into base materials. Following a similar approach to [83], this toy example is clustered into 3 base materials. In the next step, an initial BRDF slice is sampled for each base material.

we call BRDF descriptor, by quantizing the half angle space into Q' equally sized bins over $\theta_h = [0 \dots \pi/2]$ and taking a weighted sum of all color observations that belong to the same bin. In our experiments, $Q' = 3$. Consequently the BRDF descriptor's bins are transformed to LAB color space. In order to partition these BRDF descriptors into K base materials we use the method proposed by [147]. The way to determine the total number of base materials will be discussed in Sec. 3.5.5. For the toy example depicted in Fig. 3.6, this approach results in 3 base materials.

When the number of base materials have been determined and each 3D point of the mesh has been assigned to one cluster, we re-sample the BRDF slice for each base material k by using all of its assigned measurement pairs: $\{\mathbf{c}, \theta_h\}_{ij} \mid i \in k$. Again, we quantize the half angle space into Q equally sized bins (in our experiments $Q = 30$). After having assigned all RGB-triples to one of the bins, the mean value of each bin is computed to form the initial BRDF slices that can be seen in Fig. 3.6. In a later step, the BRDF $\mathbf{B}$ of each point is estimated

to be a weighted sum of K BRDF slices $\mathbf{B}_k$, when K different base materials have been found:

$$\mathbf{B}(\theta_h) = \sum_{k=1}^{K} w_k \mathbf{B}_k(\theta_h). \tag{3.4}$$

3.5.5 Determining the Number of Base Materials

In order to cluster the BRDF descriptors into K base materials of similar reflectance, we use weighted k-means. For the moment, assume that K is known. Starting from a random labeling, we compute each cluster center as the mean of all its BRDF descriptors, discarding any missing values. For element q in cluster $\mathbf{c}^{(k)}$, we have:

$$\mathbf{c}_q^{(k)} = \frac{1}{\sum_{\mathbf{d}\in\mathbf{c}^{(k)}} g(\mathbf{d}_q)} \sum_{\mathbf{d}\in\mathbf{c}^{(k)}} g(\mathbf{d}_q)\mathbf{d}_q, \tag{3.5}$$

where $g(\mathbf{d}_q)$ is an indicator function, returning 1 if element q exists in BRDF descriptor $\mathbf{d}$ and 0 otherwise. The dissimilarity between BRDF descriptor $\mathbf{d}$ and cluster $\mathbf{c}^{(k)}$ is computed as the mean squared euclidean distance between their overlapping elements:

$$dist(\mathbf{d}, \mathbf{c}^{(k)}) = \frac{1}{\sum_q [g(\mathbf{d}_q) g(\mathbf{c}_q^{(k)})]} \sum_{q=1}^{Q} g(\mathbf{d}_q) g(\mathbf{c}_q^{(k)}) (\mathbf{d}_q - \mathbf{c}_q^{(k)})^2. \tag{3.6}$$

As $\mathbf{d}$ is sparse, Eq. 3.6 is not suitable for directly computing the dissimilarity between two BRDF descriptors. However, since weighted k-means only requires the distance between $\mathbf{d}$ and the denser $\mathbf{c}^{(k)}$, this measure is applicable as long as the span of $\mathbf{c}^{(k)}$ is large enough to include $\mathbf{d}$. To reduce the sparseness of $\mathbf{d}$, we use a relatively small dimensionality for the BRDF descriptor ($Q = 3$ in our experiments). As our results show, this low angular resolution still allows for an efficient clustering of the reflectance space.

Since the output of k-means depends on the initial labeling, we run the clustering 10 times and select the labeling with the lowest error, measured as the accumulated distance between the BRDF descriptors and their corresponding cluster centers:

$$E = \sum_d dist(\mathbf{d}, \mathbf{c}_d^{(k)}) \tag{3.7}$$

Note that the clusters themselves are not used as base materials BRDFs due to their low angular resolution ($Q = 3$). Instead, we use the resulting labeling to sample high-resolution BRDF slices ($Q = 30$), as explained in Sec. 3.5.4.

To determine the total number of base materials, we start with a single material ($k = 1$) and increase k until convergence, i.e. until $E_{k+1}/E_k > 0.9$ using the error measure from Eq. 3.7.

3.6 Reflectance and Geometry Refinement

Due to lack of structure in the initial mesh, specular reflections and noise, there are inaccuracies on the initial 3D points and normals and the initial 1D BRDFs may underestimate possible reflections. In a next step, we would like to refine the base material (1D) BRDFs, photometric normals[4], material weights and 3D points positions, such that the estimated reflectance for each point in each image fits the observations. Obviously, these are a lot of parameters to take into account. In order to control this process, we propose a new optimization method that is alternating between optimizing 1D BRDFs, normals, per point weights for base materials, and finally 3D points positions.

3.6.1 Optimizing Base Materials BRDFs, Photometric Normals and Material Weights

Given an initialization for 3D points $\mathbf{P}$ and normals $\mathbf{n}$ from SfM + MvS (see Sec. 3.4), and an initialization for material weights w and base material (1D) BRDFs $\mathbf{B}$ from base material clustering (see Sec. 3.5), we aim to adjust the state:

$$x = \{\mathbf{B}_1, .., \mathbf{B}_k, \mathbf{n}_1, .., \mathbf{n}_N, w_{11}, .., w_{kN}, \mathbf{P}_1, .., \mathbf{P}_N\}. \tag{3.8}$$

To this end, we minimize a cost function with respect to constraints imposed on the base materials BRDFs, photometric normals and material weights:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} \zeta\left(i,j\right) d_{LAB}\left(\tilde{c}_{ij}\left(x\right), c_{ij}\right) \text{w.r.t. } \{E_B, E_n, E_w\}, \tag{3.9}$$

with d, the difference in LAB color space between modeled $\tilde{c}_{ij}$ (from Eq.3.1) and observed measurements c_{ij} and $\zeta\left(i,j\right)$ the visibility function that equates to one if the point $\mathbf{P}_i$ is visible in image j, and to zero otherwise. Below we describe the different constraints that are taken into account.

[4]Inspired by the works of [68], [152] and [43] we decouple between the mesh normals and the direction vectors estimated by PS. As established in previous works, we also call the latter photometric normals and assume that they are independent from the mesh normals. In a later step we are going to optimize the 3D points positions so that the mesh normals are coinciding with the photometric ones.

Base materials BRDFs The BRDF constraints E_B will be explained shortly after describing why such a BRDF slice refinement is crucial. The true HDR BRDF slice might very well extend beyond the saturation point imposed by the LDR image measurements (which is 1 in our case), especially when close to a specular highlight ($\theta_h \approx 0°$). Thus, saturated image measurements only provide limited information regarding the true form of the BRDF slice. For all we know, the true BRDF values can be anywhere above 1. Therefore d may prevent the BRDF to exceed 1 if we attempt to fit these measurements directly. At the other end of the BRDF slice ($\theta_h \approx 90°$), the quality of the image measurements deteriorates as the view-to-surface angle gets about parallel to the surface tangent. Therefore, we discard any measurements for which $\cos\theta_h < 0.15$. In summary, our BRDF slice has two loose ends that we need to somehow tie up.

Working in LAB space, we encourage a low second derivative of the luminance component L (BRDF smoothness), and a low first derivative of the chromaticity components a^* and b^* (color constancy). The first constraint ensures that the BRDF slice will be a smooth function where as the second reflects the fact that most commonly encountered BRDFs show only a minor change in color, but may exhibit large variations in intensity,

$$E_B = \frac{1}{Q}\sum_{q=1}^{Q}\left(\nabla^2\mathbf{B}_L^2(q) + \nabla\mathbf{B}_{a^*b^*}^2(q)\right). \tag{3.10}$$

Of course one might argue that apart from the surface albedo, the reflective color and incident light color might be quite different. However, the proposed constraints (especially the color constancy constraint) control erratic introduction of false colors nearby the specular peaks, which can be very disturbing in the final renderings. Specifically, we have experimentally found that when not setting the color constancy constraint the RGB bands of the BRDF slice can arbitrarily cross each other, especially above the saturation limit imposed by the LDR input images where we have no measurements to fit. This results in erroneous colors in the specular reflections, as mentioned above, which are clearly visible when rendering the object under HDR lighting. The proposed color constancy constraint helps alleviate this problem, allowing for more realistic renderings of the scanned model.

Unlike other approaches that try to predict the BRDF behavior in regions with low or no measurements by encouraging similarity with BRDFs sampled in the real world, particularly the MERL BRDF database [99], we have found that our constraints suffice to arrive at a photo-realistic BRDF estimation without sacrificing generality. In that regard, we remind that in this chapter our overall goal is not only to arrive at an improved geometry and normal distribution, but also a weighted BRDF representation that can be used to create photo-realistic renderings of the resulting 3D model.

Photometric normals Although calculating photometric normals for 3D points reflecting texture-rich regions of the illumination is straightforward, their estimation may be problematic for dark areas with little to no measurements. This is not so uncommon due to the fact that our setup, unlike traditional PS approaches, has only a single fixed lighting direction for every camera viewpoint. In order to estimate a proper photometric normal for such 3D points we rely on their neighbors. Specifically, we encourage the local surface curvature to be constant. This is approximated as the angular difference between the normalized photometric normal of each 3D point and the normalized mean photometric normal of all its neighbors,

$$E_n = \frac{1}{N}\sum_{i=1}^{N}\left(\arccos(\mathbf{n}_i\mathbf{n}_m)\right)^2 \text{ with } \mathbf{n}_m = \frac{1}{V}\sum_{v=neighbors(i)}^{V}\mathbf{n}_v. \tag{3.11}$$

Material weights In Sec. 3.5 we described the concept of base materials and explained how to subdivide the scanned object into K clusters of similar reflectance (BRDF slices) and assign one cluster to each 3D point. However, this initially assigned hard-labeling makes the object look artificial when rendered. This is mainly due to the fact that most real surfaces exhibit large variations in their reflective behavior even within the same base material (e.g. texture). In many cases, these variations can not be properly explained even by a single base material but require the combination of multiple base materials. To account for such effects, we express the individual BRDF B of each 3D point as a weighted sum of the K recovered BRDF slices B_k that represent the base materials (see Eq. 3.7). During optimization though these weights can take arbitrary values if left unconstrained. To remedy this situation, we favor the presence of positive weights by heavily penalizing any negative values,

$$E_w = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K}(\epsilon_w)^2\right) \text{ with } \epsilon_w = \begin{cases} 0 & w_k \geq 0. \\ \lambda_w w_k & w_k < 0. \end{cases} \tag{3.12}$$

The parameter λ_w controls the strength of the constraint. In our experiments $\lambda_w = 10$. To summarize, the material weights constraint nicely accounts for both concerns: (1) allows for more generality, e.g. a 3D point that was incorrectly assigned to a cluster initially can shift to another more appropriate cluster, and (2) prohibits the presence of implausible solutions by favoring positive values.

3.6.2 Optimizing 3D Point Positions

Having adjusted the photometric normals, the new estimates can be used to improve the geometry of the scanned object. In particular, we would like to

optimize the xyz-coordinates of every 3D point such that the mesh normals coincide with the newly estimated photometric normals. For this task we rely on the approach proposed by Nehab *et al.* [104]. They are optimizing for 3 independent coordinates using the following objective function,

$$\begin{aligned} &\text{argmin } \lambda_p E^p + (1-\lambda_p)E^n, \text{ with:} \\ &E^n = \sum_{i=1}^{N} \sum_{j,k \in \Omega_i} [\mathbf{n}_i^m \cdot (\mathbf{P}_j - \mathbf{P}_k)]^2, \\ &E^p = \sum_{i=1}^{N} [\mathbf{T}_i \cdot (\mathbf{P}_i - \mathbf{P}_i^m)]^2, \text{ and} \\ &\mathbf{T}_i = \alpha_p \mathbf{n}_i^m \mathbf{n}_i^{mT} + \beta_p (\mathbf{I} - \mathbf{n}_i^m \mathbf{n}_i^{mT}). \end{aligned} \tag{3.13}$$

The first term E^n, called *normal error* in their paper, allows 3D points to move such that, the tangent plane formed by the neighbors Ω_i of each 3D point $\mathbf{P}_i$, shows a normal close to the estimated photometric normal. To fully understand this term, consider the polygon formed by the neighbors of each vertex as an approximation of its tangent plane. Then, each edge in each polygon, $\mathbf{P}_j - \mathbf{P}_k$, should become perpendicular to the estimated photometric normal at the central vertex, $\mathbf{n}_i^m$. The second term E^p, referred to as *position error* in their paper, helps to prevent self-intersections in the resulting 3D model. It does so, by favoring each vertex $\mathbf{P}_i$ to move along the direction of its photometric normal, $\mathbf{n}_i^m \mathbf{n}_i^{mT}$, and by penalizing its motion across the tangent plane which is perpendicular to its photometric normal, $\mathbf{I} - \mathbf{n}_i^m \mathbf{n}_i^{mT}$. The parameters α_p and βp control the penalty of the movement along the photometric normal and across the tangent plane perpendicular to the photometric normal, respectively. As such, choosing a relatively low value for α_p and a relatively high value for βp helps in achieving the desired result. Notice, though, that these parameters are dependent by the total number of 3D points N. In our experiments, $\alpha_p = 1 \times 10^5 / N$ and $\beta_p = 1 \times 10^3 \cdot \alpha_p$. Finally, the parameter λ_p controls how much influence the position and normal error have in the optimization. In our experiments, $\lambda_p = 0.1$.

Minimizing the objective function directly, though, by using energy-based minimization techniques, as done in [104], has two inherent limitations: (1) an optimal solution is hard to be reached since the system is prone to get stuck in local minima, and (2) as the number of 3D points increases the system becomes intractable due to the large amount of memory that it requires. To deal with these limitations, we re-formulate the problem in a least squares sense to arrive at a closed form solution. Notice that before squaring, the equations for the error terms are linear in the 3D points' coordinates we are solving for. Therefore,

our problem can be re-written as an over-constrained linear system of equations, which can consequently be solved using least squares.

Each 3D point centered at a polygon with v edges generates $3 + v$ equations: 3 for the position error, $\lambda_p \mathbf{T}_i \cdot \mathbf{P}_i = \lambda_p \mathbf{T}_i \mathbf{P}_i^m$[5], and v for the normal error, $(1 - \lambda_p)\mathbf{n}_i^m \cdot \mathbf{P}_j - (1 - \lambda_p)\mathbf{n}_i^m \cdot \mathbf{P}_k = 0$, $\forall j, k \in \Omega_i$. As such, the linear system of equations can be summarized as:

$$\begin{bmatrix} \lambda_p \mathbf{T}_i & 0 & 0 & \dots \\ 0 & (1-\lambda_p)\mathbf{n}_i^m & -(1-\lambda_p)\mathbf{n}_i^m & \dots \\ & \dots\dots\dots\dots\dots & & \\ & \dots\dots\dots\dots\dots & & \end{bmatrix} \cdot \begin{bmatrix} \mathbf{P}_i \\ \mathbf{P}_j \\ \mathbf{P}_k \\ \dots \end{bmatrix} = \begin{bmatrix} \lambda_p \mathbf{T}_i \mathbf{P}_i^m \\ 0 \\ \dots \\ \dots \end{bmatrix} \tag{3.14}$$

For N 3D points the total number of equations becomes very high, $N \cdot (3 + v)$, but since the matrix is very sparse, having at most 2 non-zero entries per row, it can efficiently be solved using the Conjugated Gradient algorithm [119] for sparse linear least squares systems.

3.6.3 Minimization Details

We minimize Eq. 3.9 w.r.t. the different constraints E_B, E_n, E_w using non-linear least squares minimization [1]. Optimizing multiple parameters at once is inefficient for two reasons: (1) the system is unstable, getting more easily stuck at local minima and (2) a full global refinement is computationally very expensive. From our experience, in such multi-parameter optimization problems it is important to optimize each class of parameters independently and constraint the space of possible solutions in order to arrive at plausible results.

Motivated by these observations, we optimize in 4 discrete steps. For the current 3D points positions: (1) We refine the base materials BRDF slices $\mathbf{B}_k$ for a selection of measurements (10% of their total number randomly chosen in our experiments) where $i \in k$ for each cluster individually keeping the material weights w_i constant. (2) We fix the base materials BRDFs $\mathbf{B}_k$, and refine the photometric normals $\mathbf{n}_i$ for compact groups of 3D points (max number of 3D points per group is 1×10^5). (3) We optimize the material weights w_i for each 3D point individually. (4) Finally, having calculated new photometric normals we use these estimates to update the 3D points positions $\mathbf{P}_i$ by solving the sparse linear system of Eq. 3.14 using least squares.

[5]Remember that $\mathbf{T}_i$ is a 3×3 matrix.

3.7 Results

The proposed pipeline has been tested on various, synthetic and real, challenging examples, both in terms of recovering 3D shape as well as reflectance. Specifically, we first quantitatively evaluated our approach on synthetic 3D models using specular MERL BRDF samples and compared our method with existing approaches [104, 43]. Starting from a varying number of synthetically rendered images we added different types and levels of noise or smoothing and verified the ability of each method to recover the ground truth geometry of the object. In a second set of experiments, we investigate the performance of our algorithm in estimating reflectance. To this end, for every MERL BRDF we rendered synthetic LDR and HDR images of blobs, which we then corrupted with noise, and verified the effectiveness of our pipeline in recovering the ground truth reflectance. In a third set of experiments, we tested the sensitivity of our setup with respect to the light's deviation from the camera as well as in the presence of image noise. Finally, we show 3D reconstruction and rendering results on numerous real-world objects, varying in several aspects[6] and starting from different types of mesh initializations.

3.7.1 3D Shape Evaluation from Synthetic Data

In this section, we quantitatively evaluate the performance of our pipeline in recovering an object's 3D shape. For this task, we used 3 synthetic models from the Stanford 3D scanning repository with an increasing level of geometric detail, *Armadillo* (350K faces), *Dragon* (850K faces) and *Happy-Buddha* (1100K faces) respectively. For each model we rendered synthetic 2048×2048 images using a different specular MERL BRDF sample, *red-specular-plastic* for *Armadillo*, *silver-paint* for *Dragon* and *gold-metallic-paint* for *Happy-Buddha*. We specifically chose highly reflective samples to show the effectiveness of our method in handling such cases. We created in total 90 images per 3D model from different viewpoints that are uniformly sampled around the object. For every viewpoint the light is co-located with the camera's center. It is assumed that all the different approaches in the quantitative evaluation have access to the true camera intrinsic and extrinsic parameters and light positions. Since the chosen 3D models vary in their size, we center and scale them so that the radius of their tightest bounding sphere is unit. This is also done for numerical consistency in the quantitative results. Next, we added different types and levels of noise, decimation, smoothing, to simulate errors and irregularities in real-world data.

[6] The number of base materials, the presence or not of texture, the level of geometric detail, the surface reflectance characteristics, etc.

Firstly, we perturb the original 3D model by adding random vertex displacements and we consequently apply the Taubin smoothing operator [146]. This is to simulate the mesh initializations that are recovered from our SfM pipeline. In total, we generated 3 levels of mesh perturbation with an increasing number of noise (from level 1 having the lowest error till level 3 that has the highest error). Secondly, we decimate the original 3D model by using a mesh simplification technique [65] and we then apply Gaussian smoothing to simplify the high-frequency details. This is to mimic the mesh initializations that come from visual-hull + PSR or SfM (sparse point-cloud) + PSR. In total, for each 3D model we generate meshes with approximately 1/2, 1/4, 1/8 of the original number of faces before applying 5 iterations of Gaussian smoothing. In both cases, the ability of each approach to recover the ground truth geometry is also verified with respect to the number of input images. Specifically, we ran each experiment with 90, 60 and 45 images, respectively.

To quantify the geometric similarity of the recovered 3D model R with respect to the ground truth G, we used the *accuracy* (how close R is to G) and *completeness* (how much of G is modeled by R) metrics, as originally defined by the Middlebury multi-view stereo benchmark [133]. In particular, for measuring the accuracy the authors compute the asymmetric distances $dist_{R \to G}$ between the points in R and the nearest points in G. Then, the accuracy is simply the distance $d \in dist_{R \to G}$ such that $X\%$ of the points on R are within distance d of G. Similarly, to measure the completeness the authors compute the asymmetric distances $dist_{G \to R}$ from G to R, i.e. the opposite of what they did for measuring accuracy. Now, the completeness is defined as the percentage of vertices where $dist_{G \to R}$ is less than a threshold $dist_{th}$. In our experiments, $X = 90$ and $dist_{th} = 0.01$. Notice that lower accuracy and higher completeness means better results.

As already explained, having only a consumer camera with flash as our recording setup restricts the applicability of approaches, like [62, 124, 169, 110], that either require custom-built designs (cameras, tripods), specific equipment (light tubes, BRDF chart), multiple calibrated lights or additional information (segmentation masks, environment maps), etc. One might argue, though, that a multi-view PS approach can still be applicable, even if only a single light is available per viewpoint. As such, we compared our approach with the method of Hernandez *et al.* [43]. We also included the original method of Nehab *et al.* [104] in our evaluation. To achieve a fair comparison we used outliers handling in the latter approaches too, since they typically assume a Lambertian reflectance model. Tbl. 3.1 summarizes the quantitative results on the synthetic data.

For the mesh perturbation experiment, we observe that overall our method consistently performs better than the approaches of Hernandez *et al.* [43] and Nehab *et al.* [104] in both accuracy and completeness metrics. This is to be

Table 3.1: 3D shape evaluation from synthetic data. Each method refines perturbed or decimated meshes for a different number of input images, and the results are assessed in comparison with the ground truth. Each cell shows the *accuracy* (10^{-4}) and *completeness* (%) metrics [133], for the mesh perturbation and resolution experiments. See also Sec. 3.7.1.

Armadillo (Mesh Resolution: 350K)

Mesh Perturbation	Level 1			Level 2			Level 3		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	3.43, 98.6	3.44, 98.6	3.47, 98.5	3.46, 98.6	3.47, 98.5	3.49, 98.4	3.47, 98.4	3.48, 98.4	3.50, 98.3
Hernandez *et al.* [43]	3.13, 99.4	3.16, 99.4	3.23, 99.2	3.15, 99.4	3.20, 99.2	3.26, 99.0	3.16, 99.4	3.22, 99.3	3.27, 99.2
Our method	**2.74**, **99.9**	**2.77**, **99.9**	**2.80**, **99.8**	**2.76**, **99.9**	**2.79**, **99.8**	**2.81**, **99.8**	**2.80**, **99.8**	**2.84**, **99.8**	**2.91**, **99.7**
Mesh Decimation	150K			75K			35K		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	2.98, 98.8	2.99, 98.7	2.99, 98.7	3.98, 96.0	3.99, 96.0	4.00, 96.0	5.80, 87.2	5.80, 87.0	5.86, 86.3
Hernandez *et al.* [43]	2.92, 98.9	2.94, 98.9	2.95, 98.8	3.98, 96.0	4.00, 96.0	4.05, 95.8	5.70, 87.5	5.73, 87.4	5.80, **87.0**
Our method	**2.55**, **99.4**	**2.65**, **99.3**	**2.69**, **99.2**	**3.72**, **96.9**	**3.75**, **96.8**	**3.78**, **96.8**	**5.59**, **88.0**	**5.63**, **87.6**	**5.71**, **87.0**

Dragon (Mesh Resolution: 850K)

Mesh Perturbation	Level 1			Level 2			Level 3		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	4.44, 93.8	4.46, 93.5	4.58, 92.8	4.51, 93.7	4.55, 93.5	4.57, 93.4	4.50, 94.0	4.53, 93.8	4.57, 93.5
Hernandez *et al.* [43]	3.95, 94.4	3.96, 94.3	4.00, 93.5	4.03, 96.3	4.05, 96.2	4.08, 95.8	4.03, 96.6	4.09, 96.4	4.15, 96.0
Our method	**2.87**, **99.8**	**2.92**, **99.7**	**3.04**, **99.6**	**2.95**, **99.7**	**2.98**, **99.6**	**3.05**, **99.5**	**3.08**, **99.6**	**3.10**, **99.6**	**3.17**, **99.5**
Mesh Decimation	400K			200K			100K		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	3.55, 97.5	3.57, 97.5	3.80, 96.8	3.61, 97.3	3.61, 97.3	3.81, 96.6	4.06, 95.5	4.08, 95.4	4.23, 94.8
Hernandez *et al.* [43]	3.53, 98.2	3.56, 98.1	3.57, 98.0	3.56, 98.1	3.60, 97.9	3.65, 97.6	3.95, 96.4	3.97, 96.2	4.03, 95.7
Our method	**1.55**, **99.9**	**1.60**, **99.9**	**1.81**, **99.9**	**2.02**, **99.9**	**2.06**, **99.9**	**2.20**, **99.8**	**2.99**, **98.8**	**3.03**, **98.7**	**3.14**, **98.5**

Happy-Buddha (Mesh Resolution: 1100K)

Mesh Perturbation	Level 1			Level 2			Level 3		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	4.36, 95.5	4.35, 95.0	4.27, 94.7	4.56, 93.7	4.58, 93.1	4.60, 92.9	4.70, 92.7	4.71, 92.2	4.72, 91.9
Hernandez *et al.* [43]	3.65, 97.4	3.68, 97.2	3.70, 97.0	3.69, 97.6	3.70, 97.6	3.73, 97.4	3.80, 97.4	3.83, 97.3	3.86, 97.1
Our method	**2.73**, **99.9**	**2.81**, **99.7**	**2.91**, **99.6**	**2.74**, **99.7**	**2.80**, **99.6**	**2.90**, **99.5**	**2.85**, **99.5**	**2.92**, **99.5**	**3.00**, **99.4**
Mesh Decimation	500K			250K			125K		
Number of Images	90	60	45	90	60	45	90	60	45
Nehab *et al.* [104]	3.73, 97.0	3.74, 96.9	3.76, 96.9	3.95, 95.9	3.96, 95.9	3.96, 95.8	4.71, 91.8	4.72, 91.5	4.75, 91.1
Hernandez *et al.* [43]	3.10, 98.7	3.12, 98.6	3.19, 98.5	3.38, 98.0	3.41, 97.9	3.49, 97.6	4.30, 93.1	4.34, 92.7	4.39, 92.4
Our method	**1.91**, **99.9**	**1.98**, **99.9**	**2.14**, **99.8**	**2.71**, **99.2**	**2.79**, **99.2**	**2.90**, **99.1**	**3.99**, **95.1**	**4.08**, **94.7**	**4.19**, **94.1**

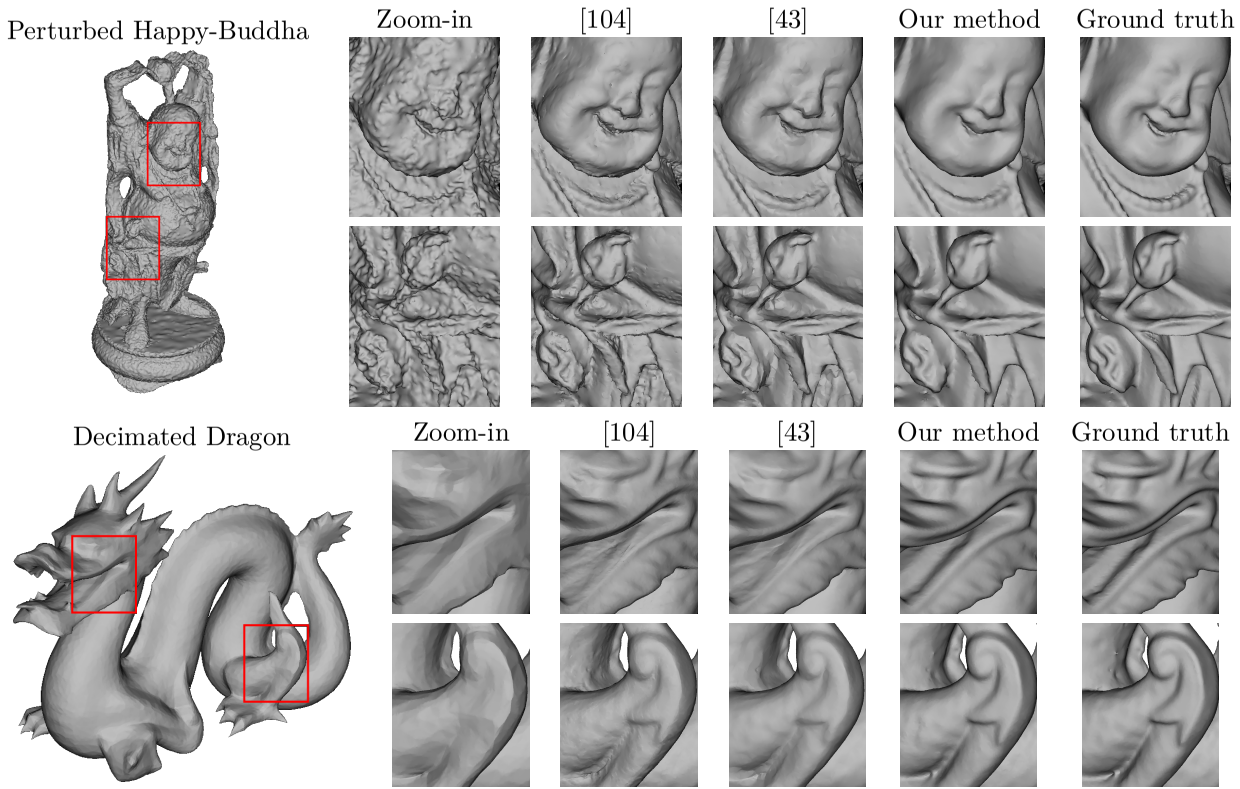


Figure 3.7: A perturbed *Happy-Buddha* (Tbl. 3.1, mesh perturbation: level 3, number of images: 60) and a decimated *Dragon* (Tbl. 3.1, mesh decimation: 100K, number of images: 60) model and refinement results for the three approaches.

expected as our approach naturally handles specularities to arrive at better photometric normal estimates, instead of discarding them as outliers [104, 43], which in turn guides the movement of the 3D points closer to their ground truth positions. Furthermore, as specular reflections pose strong constraints on the direction of the photometric normals [24, 167], discarding them and assuming that there is an approximately Lambertian behaviour for at least a subset of the viewing/lighting combinations [104, 43] is not enough for our camera/flash setup. The latter is visible in the quantitative results of Tbl. 3.1 especially as the number of input images goes from 90 to 45, where we see that the performance of our approach remains reasonably stable, mainly for the completeness metric, whereas that of [104, 43] degrades. Another observation is that the accuracy and completeness of Nehab *et al.* [104] is generally lower than ours but also lower than Hernandez *et al.* [43] since their approach is prone to mesh flipping and overlapping triangle effects. For relatively simple meshes, like *Armadillo*, the performance difference between the three approaches is minimal but as we move to examples with more geometric details, like *Dragon* and *Happy-Buddha*, our method clearly comes first by a large margin. This improvement is not only numerical though. As shown in the upper part of Fig. 3.7 that visualizes the refined meshes for *Happy-Buddha*, our method faithfully reconstructs the fine surface details such as the facial features, the necklace and the flower on the model.

For the mesh resolution experiment, the general observations coincide with the mesh perturbation experiment. In particular, our method outperforms [104, 43] for both metrics, possibly indicating some sort of robustness to worse initializations, like the ones from visual-hull + PSR or SfM (sparse point-cloud) + PSR. The only exception being that of the highest decimation for *Armadillo* and *Happy-Buddha*, where all methods fail to reconstruct the fine details. The latter is an inherent limitation of all PS-based approaches since for very approximate initializations the initial geometry projects to the wrong pixel positions, especially as the view direction deviates from the surface normal, accumulating many erroneous measurements and as a result making the convergence difficult even when outliers handling is used. As the number of input images is reduced we observe that the performance of our approach remains reasonably stable in contrast to [104, 43], mainly for the completeness metric, due to inclusion of specular measurements in the photometric normals estimation as also explained above. Overall, since we rely on a least-squares closed form solution for optimizing the 3D points positions compared to energy-based minimization techniques of [104, 43], it is more likely to arrive closer to the ground truth positions. As such, our method recovers fine geometric details even from lower resolution initial meshes, as can be seen for the mouth and leg of *Dragon* in the bottom part of Fig. 3.7.

3.7.2 Surface Reflectance Recovery from Synthetic Data

So far, we have assessed our approach with respect to its ability to recover 3D shape. In this section, we investigate the performance of our method in estimating surface reflectance. A simple way to evaluate the recovered reflectance is by setting up a series of synthetic experiments under point light illumination. We begin by considering an object, in particular a blob. For each one of the 100 MERL BRDFs we render synthetic images depicting the blob with the corresponding MERL BRDF under point lighting. We assume the light is aligned with the camera to mimic our camera/flash setup. In total, we render 30 512×512 images from different viewpoints around the blob and keep two variants for each image, a HDR and a LDR one, required for the following analysis. Next, we perturb the blob by adding random vertex displacements and applying Taubin smoothing, similar to the level 3 mesh perturbation experiment of Sec. 3.7.1. Finally, starting from the perturbed mesh we run our full pipeline and compare the recovered with the ground truth BRDFs.

To evaluate our BRDF estimates, we use the log-space RMSE between the recovered and ground truth non-parametric BRDF slices, which is the established protocol [90, 93]. The upper part of Fig. 3.8 summarizes the quantitative results for all MERL BRDFs when HDR and LDR images are used as input to our method. The results are arranged by descending log-space RMSE order. Overall, we observe that for 90% of MERL samples the non-parametric BRDF slice is recovered with less than 0.4 log-space RMSE for both HDR and LDR inputs. This indicates that our approach can faithfully recover the BRDF slice even from difficult initializations (*i.e.* mesh perturbation: level 3), which can also be seen in the bottom part of Fig. 3.8 that visualizes renderings of the different BRDF slices (ground truth, recovered from HDR and LDR inputs) for 4 MERL samples in blobs. In general, the renderings look identical and the only source of error is found in the shape or size of the specular peak. Another observation is that specular BRDFs are more difficult to estimate and they result in higher error. This is likely because the specular BRDFs have a greater degree of variation. Regarding the method's efficiency to recover the BRDF slice when HDR or LDR inputs are used, Fig. 3.8 shows that the log-space RMSE is consistently lower for HDR inputs. This is to be expected since recovering BRDF information from LDR images is a significantly harder problem compared to the HDR case. Nevertheless, our method can still estimate the BRDF slice from LDR inputs up to an unknown specular peak magnitude. Note that this is not straightforward; due to saturation point imposed by the LDR image measurements we have no way to know the true magnitude of the specular peak (see Sec. 3.6.1). As seen in the bottom part of Fig. 3.8, however, the latter has minimal impact on the final rendering.

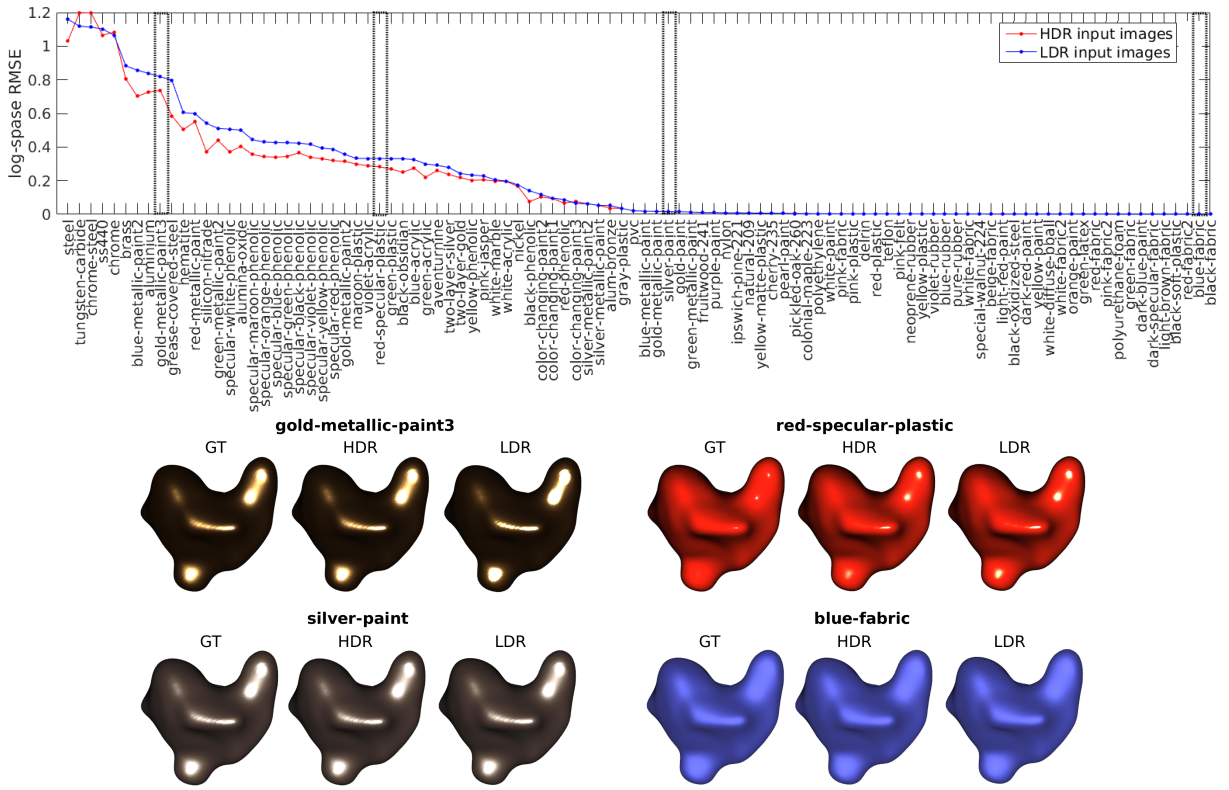


Figure 3.8: Surface reflectance recovery from synthetic data. On the upper part, this figure shows the log-space RMSE of recovered BRDF slices for the 100 MERL samples from HDR and LDR inputs. On the bottom part, we highlight several example results to illustrate the log-space RMSE values correspond to perceptual accuracy. See also Sec. 3.7.2.

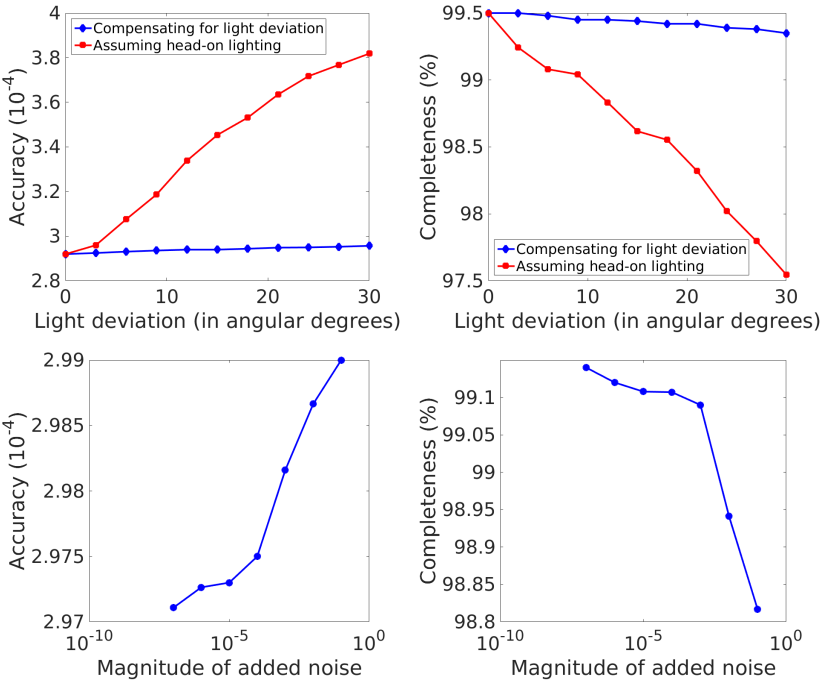


Figure 3.9: Sensitivity to light deviation and image noise. On the upper part, this figure shows the accuracy and completeness metrics for the light deviation experiment when assuming head-on lighting and when compensating for the actual light position. On the bottom part, we visualize the same metrics for the image noise experiment. For more details see Sec. 3.7.3.

3.7.3 Sensitivity to Light Deviation and Image Noise

In the following set of synthetic experiments we would like to evaluate the sensitivity of our pipeline with respect to aspects related to the scanning procedure. First, we test the system's ability to recover geometry when the position of the light, in our case the flash light, deviates further from the position of the camera. Second, we evaluate the method's performance in the presence of image noise. The next paragraphs provide more details for each experiment.

For the light deviation experiment, we assume the blob of Sec. 3.7.2 rendered with the MERL material *blue-metallic-paint2* from 30 different but fixed viewpoints around the blob. In total, we render 11 series of 30 512×512 images where the position of the light deviates from the position of the camera between

0 and 30 angular degrees with a incremental step of 3 angular degrees each time. Next, we perturb the blob by adding random vertex displacements and applying Taubin smoothing, similar to the level 3 mesh perturbation experiment of Sec. 3.7.1. For each of the 11 different camera/light configurations, starting from the perturbed mesh we run our full pipeline, up to the optimization of the 3D points positions, and estimate the accuracy and completeness as defined in the Middlebury multi-view stereo benchmark [133]. Every time we run two variants of the same experiment, one where we assume the camera and light are co-located and another one where we account for the actual light position. The first variant aims at evaluating the method's sensitivity to the assumption of head-on lighting. The accuracy and completeness graphs of these experiments are presented in the upper part of Fig. 3.9. As expected, not accounting for light deviation results in a decrease in performance with respect to both metrics, which is almost linear up to 30 angular degrees of deviation in our experiments. However, we observe that when compensating for light deviation the performance remains almost stable for both metrics.

For the image noise experiment, the setting is almost identical to the light deviation experiment described above. The only difference is that instead of deviating the light we add noise in the rendered images. In particular, we add zero-mean noise where the local variance of the noise is a function of the image intensity values, as in [87]. The magnitude of added noise variance ranges from 10^{-7} till 10^{-1} in a logarithmic scale. The recovered results are illustrated in the bottom part of Fig. 3.9 for the accuracy and completeness metrics. Going from low to high levels of added noise, and up to 10^{-3} variance, performance shows only a small decrease, but for higher magnitudes we see a drastic fall for both metrics. In general, this should not be a problem; for the majority of image sensors used in real experiments, the level of noise is typically much lower [88].

3.7.4 3D Shape and Surface Reflectance from Real Data

While synthetic evaluation is useful for generating quantitative results, our method is ultimately to be tested on real data. This section describes the results on real-life examples photographed with a DSLR camera or mobile phone with built-in flash. In both cases, we capture RAW images, from which we later create the linear images that we use during optimization. The initial geometry is generated using SfM + MvS. For each example, we first show a picture from a particular viewpoint of the original image set, on the left, and then compare the recovered geometry against the initial SfM-based geometry, on the right (Fig. 3.10). Since we do not have access to the ground truth 3D models of the scanned real-life examples, this visual comparison between the initial and recovered geometry serves as an evaluation of the estimated 3D shape. Secondly,

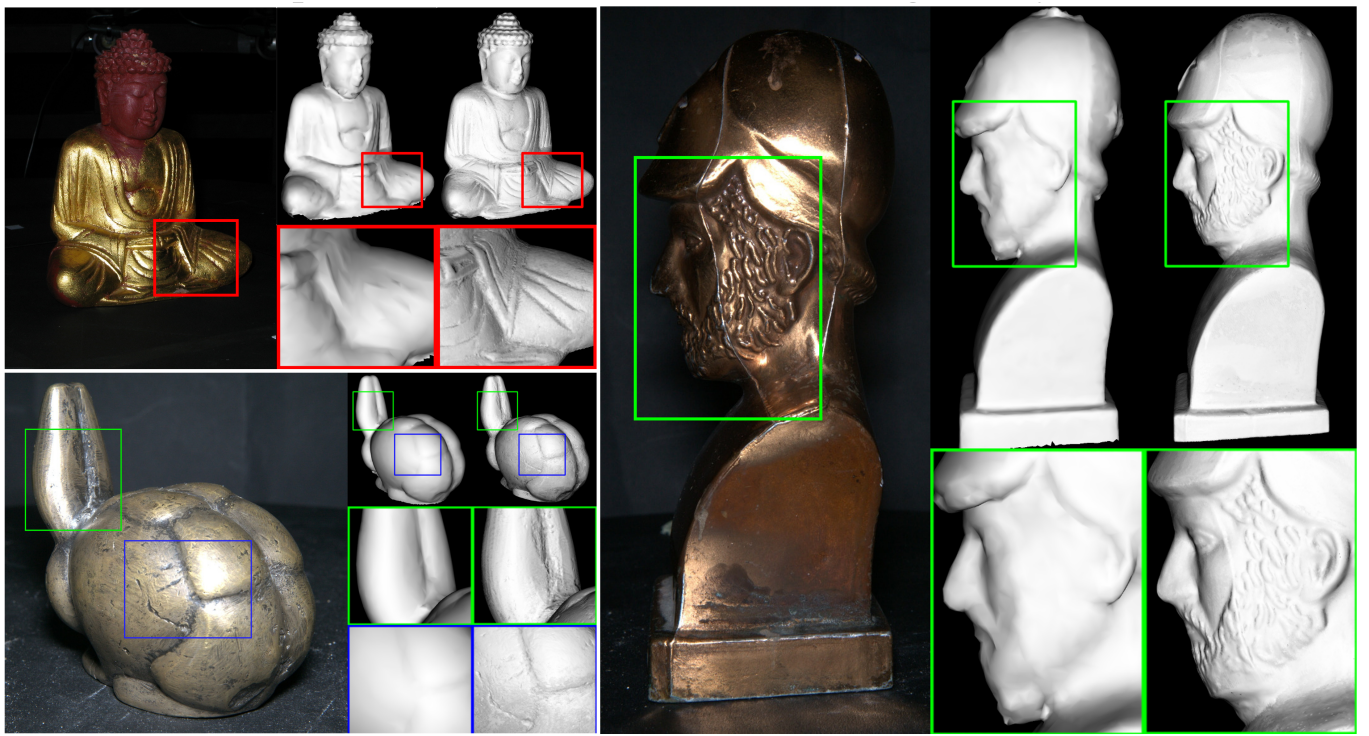


Figure 3.10: Comparison between initial and recovered geometry for *Buddha*, *Pericles* and *Rabbit* examples. For each example on the left we show an input image from the image set, in the center the initial geometry acquired using SfM + MvS and on the right the recovered geometry using our approach. The colored rectangles provide a better insight in the fine surface details and serve as an indication of the geometry improvement. See Sec. 3.7.4 for detailed comments.

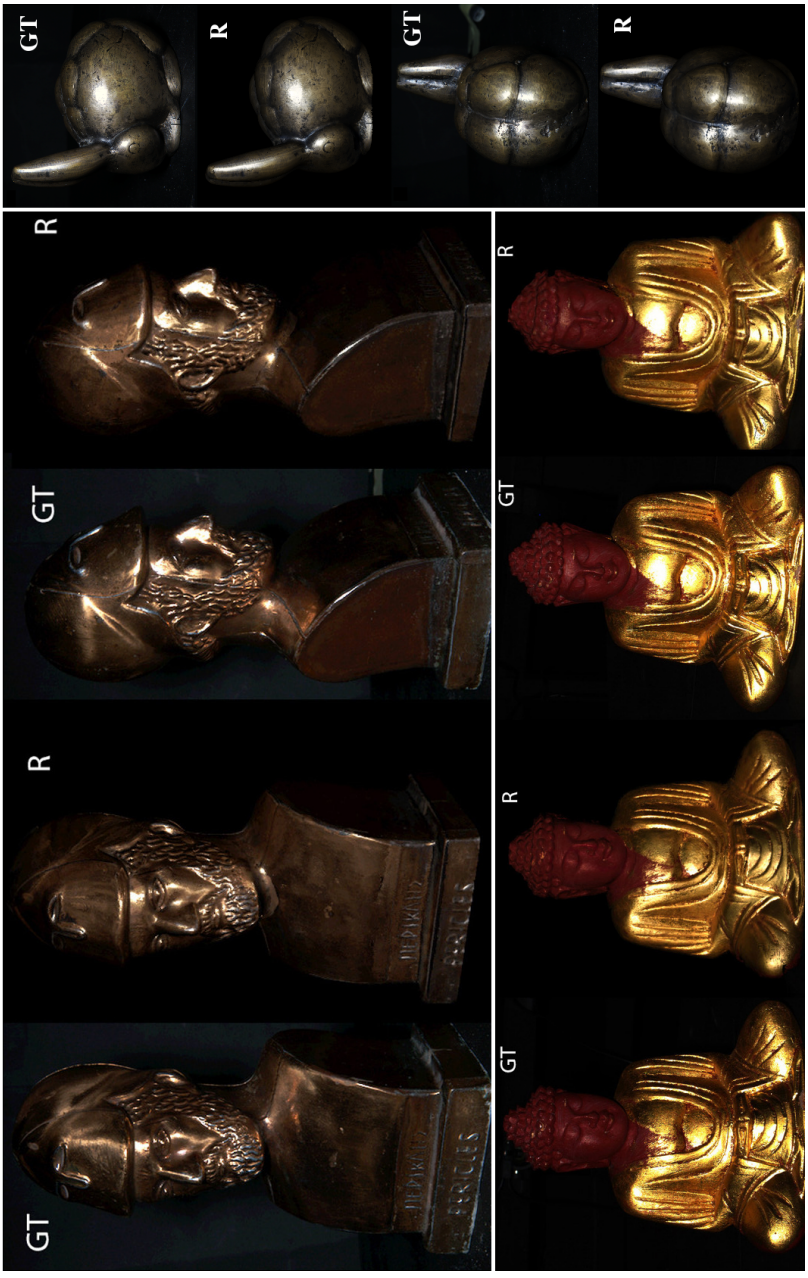


Figure 3.11: Visual comparison between ground truth and rendered linear images for the different examples of Fig. 3.10. Abbreviations: GT = ground truth image, R = rendered image.

we compare the original and rendered linear images to give a representative visualization of the recovered reflectance (Fig. 3.11). Note that these images come from viewpoints that are not included in the optimization pipeline.

Buddha A very challenging sample originating from the OBJECTS2011 data set [132]. The *Buddha* was captured with a multi-camera, multi-light dome that consists of a spherical rig of 165 flash-equipped cameras. In order to mimic our fixed-flash scenario, we singled out the images for which the flash of the capturing camera was fired. In total, we used 64 images, which is a dramatic reduction from the 27225 in the original set. Three different materials come out of the base materials clustering, a diffuse red for the head and two golden ones for the body. The object is especially challenging, because of inter-reflections, changing chromaticities and mixed materials which render the material boundaries less clear. Fig. 3.10 illustrates the difference between the initial and recovered geometry for Buddha's frontal part, where we clearly see the improvement in the right leg. Also notice how the rendered images closely match the original ones in Fig. 3.11. The recovered shape and reflectance out of this crude sample set is nevertheless convincing.

Pig-Tablet. A decorative plate with carvings showing a pig is illustrated in Fig. 3.12. The image sequence consists of a series of 10 images captured horizontally at 3 different heights (30 images in total). Two different materials have been detected in this case, a more shiny material representing the polished surface on top, and a rough version for the carvings. The initial shape from SfM is hardly able to grasp any surface details on the tablet, which do appear after the refinement process (*e.g.* the signature in the rear part of the pig). Notice how the rendered image is almost indistinguishable from the ground truth image in Fig. 3.12.

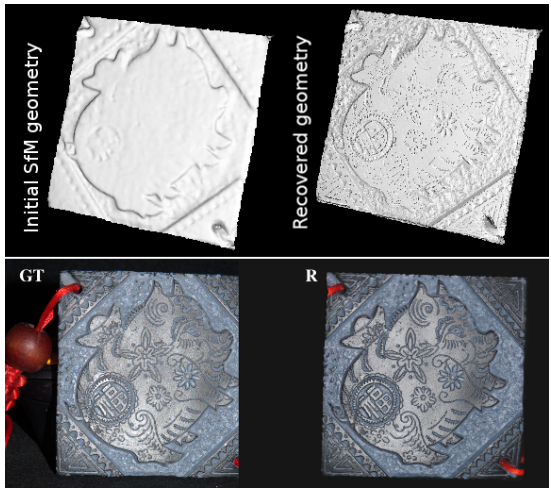


Figure 3.12: The *Pig-Tablet* example.

Pericles The *Pericles* figurine is a bronze-like statue which has a distinctive specular reflection across its surface. The image sequence consists of a series of 40 images horizontally at 3 different heights (120 images in total). The intricate reflectance allows to recover quite some details despite the crude SfM initialization, as can be seen in Fig. 3.10. For example, the face was initially very smooth lacking all the features whereas in the refined model we can see how the eyes, ears and beard are generated with high details. The recovered reflectance in Fig. 3.11 looks realistic too. Some artifacts in the upper part of the head are due to the low number of measurements recorded for this part.

Rabbit Fig. 3.10 shows an image out of a hand-held sequence of a bronze rabbit captured using an Android mobile phone with built-in flash. The refinement process gives a much better view on the overall model. In particular the gaps and cracks in the bronze are picked up. Although improved, some errors are located at the back of the ears and the top of the back, due to occlusions. Fig. 3.11 shows a visual comparison between ground truth and rendered images.

The examples show that we are able to estimate reliable 3D shape information and BRDF slices, and create virtual renderings that match the observations, despite of the complexity of the overall pipeline.

3.8 Conclusion

We have investigated the use of simple flash-based photography to capture an object's 3D shape and surface reflectance characteristics at the same time. The presented method combines the principles of SfM, MvS and PS, yet, we make sure not to use more than readily available consumer equipment, like a camera with flash or a smartphone. We experimentally validated our approach by modeling several challenging examples, both synthetic and real, ranging from diffuse till highly specular surfaces. Although acquiring accurate 3D shape and photo-realistic reflectance from such modest setup is a hard problem, our method performs better than existing hardware-dependent approaches.

From the reflectance point-of-view, in this chapter we solely focused on the estimation of BRDF slices. The question is whether one can infer a reasonable approximation of the missing, 2D or even 3D, BRDF in order to relight the object from different viewing/lighting configurations than the ones used for scanning. In Chapter 4, starting from the BRDF slices estimated with the proposed camera/flash setup we aim to provide an answer for this question. From the illumination point-of-view, the use of point lighting (*i.e.* the flash light)

typically assumes that the scanning procedure takes place inside a darkroom. To remedy this situation and relax the lighting restrictions, in Chapters 5 and 6 we investigate how to compute surface normals, parametric reflectance, and illumination under natural environmental lighting. As a final note, in the following chapters we also drop the requirement of having to capture multiple images and work with a single HDR or LDR input.

Chapter 4

Inferring Surface Reflectance

In literature the problem of estimating a full BRDF from partial observations has already been studied using either parametric or non-parametric approaches. The goal in each case is to best match this sparse set of input measurements.

In this chapter, we address a more difficult variant of this problem, *i.e.* inferring higher order reflectance information starting from the minimal input of a single BRDF slice. In the previous chapter we showed how to recover BRDF slices using a camera with flash. For the sake of generality though here we will consider the prototypical case of a homogeneous sphere, lit by a head-on light source, which only holds information about less than 0.001% of the whole BRDF domain. As such, this chapter aims to provide an answer for Research Question 2: *To what extent can we infer high-dimensional reflectance information from a single image?* To tackle this problem we propose a novel method to infer the higher dimensional properties of the material's BRDF, based on the statistical distribution of known material characteristics observed in real-life samples.

The work and analysis of this chapter correspond to:

- S. Georgoulis, V. Vanweddingen, M. Proesmans and L. Van Gool, *A Gaussian Process Latent Variable Model for BRDF Inference.* Published in IEEE International Conference on Computer Vision (ICCV) 2015.

4.1 Introduction

In the previous chapters we discussed that the appearance of an object is essentially determined by the combination of its 3D shape, its surface materials (reflectance), and the lighting environment (illumination). Producing photo-realistic renderings of an object under novel lighting is of great importance for various applications that are based on Virtual Reality (VR) or Augmented Reality (AR). For these applications one thus needs to accurately capture both the 3D shape and the surface reflectance. Yet, it is fair to say that 3D shape extraction has advanced more than the extraction of surface reflectance. In this chapter, we assume that a high-quality 3D shape of the modeled object is known in advance and we focus on precisely estimating its reflectance characteristics.

The appearance properties of opaque materials are effectively encoded by the BRDF [106], which relates incoming and outgoing directions of light transport. Specifically, this function estimates the fraction of reflected light for every pair of incoming/outgoing light directions (see Sec. 2.3.1). Typically, such BRDF has to be recorded with sophisticated hardware setups that independently drive a light source and a sensor to many different positions around the object [98, 99, 64]. These setups, however, are expensive and inaccessible to most researchers, let alone casual users. Furthermore, a dense sampling of an object's BRDF - usually only of a small planar patch - is a time-consuming process; for a sampling at an angular resolution of 1 degree more than 10^8 measurements are required [74].

In this chapter, we analyze how a complete BRDF can be inferred when only a limited number of its samples are available. In particular, we consider the use of a camera with built-in flash as in Chapter 3. In that case the viewing and lighting directions are almost identical. We assume the flash light to be dominant over other illumination in the scene and that a single image is taken. Our starting point is the prototypical case of a single image of a sphere. Unlike previous studies that consider either environment lighting [126, 91, 93] or sparse samples across the entire BRDF domain [109], in our case the coincidence of lighting and viewing directions only yields a small section of the BRDF space (see Sec. 4.2). This is a particularly difficult case compared to this considered in [126, 91, 109, 93], because not only do we have very few samples but they are also very concentrated, so in our case inferring the rest of the BRDF is more a matter of extrapolation than interpolation. We develop a solution general enough to deal with this issue, as well as to infer BRDFs of multiple dimensions. Fig. 1.2 gives a preview.

4.2 Previous Work

For the human observer, inferring reflectance information from images comes naturally. Several studies have explored how the human visual system achieves this [114, 48, 134, 158]. Fleming *et al.* [48] found that people do not need specific information about the environment to infer reflectance, but this ability declines when the environment deviates from those found in nature [48, 36].

As already mentioned in Sec. 4.1, we want to consider the special case where a camera with flash is used. In this case, considering the BRDF parameterization of Fig. 3.3, ω_i and ω_o are almost the same as long as the distance between the camera and the object is much larger than the distance between the camera and the flash, and therefore $\theta_d \simeq 0$. This yields a 1D section $f(\theta_h)$ of a 2D $f(\theta_h, \theta_d)$ or 3D BRDF $f(\theta_h, \theta_d, \phi_d)$, usually referred to as 1D BRDF or BRDF slice. Thus, one can consider BRDFs of different dimensionality, depending on the intended level of precision. The vast majority of papers in literature typically consider those dimensions to be independent, i.e. separable. In this chapter we will show that they are actually statistically dependent, indicating the relevance of higher-dimensional inference from our fringe sections. This is the first line of work that examines this dependency. No prior assumptions are made with respect to the shape of the BRDFs (e.g. number of specular lobes [21]) or the material type. In fact, as will be explained below, our method leverages the unique reflectance properties of different classes of materials (e.g. plastics, paints, etc) to arrive at better predictions. This is a core part of the training process and no user interaction is required (unlike in [21]).

Parametric approaches Parametric reflectance models have a long history in both computer vision and computer graphics. They range from ad-hoc models (e.g. Blinn-Phong [15], Lafortune [78], Ashikhmin [6], DSBRDF [107]) designed for efficiency, to physics-based derivations either based on the micro-facet theory (e.g. Ward [156], Cook-Torrance [25], Schlick [130]) or wave optics (e.g. He [59]). For a comparison of various reflectance models we refer the reader to empirical studies like [105]. There is prior work on estimating parametric reflectance models from single images, like [17, 163], but they require the functions that form the BRDF models to be defined in advance. Few methods have been designed for unknown lighting, but they also typically assume that the reflectance can be represented by a parametric BRDF model that is chosen in advance, such as Phong, Ward, or Lafortune models (e.g. [108, 162, 57]). Most recently, Lombardi and Nishino [91, 93] used a probabilistic formulation that incorporates assumptions about typical illumination environments and reflectance properties as prior distributions over latent variables to jointly estimate the most "realistic" reflectance and illumination. In general, although parametric models continue

to improve (see [107]), their usability is restricted. First of all, the reflectance model should be chosen a priori, without a guarantee that there are parameters that yield the measured data. Secondly, an error metric has to be chosen during the fitting process, not knowing which choice is optimal. Thirdly, since these models are non-linear in their parameters, the required computation is tied to the model and can not be easily transferred from one material class to the other. Furthermore, the quality of the fit is dependent on a good initial guess, and reaching a global minimum can not be guaranteed. Finally, parametric models impose restrictions on the space of materials [105, 139]. Instead, we go for a purely data-driven approach.

Semi-parametric approaches Semi-parametric models of spatially varying BRDFs for interactive editing have also been proposed (see [81]). In that case the reflectance functions are unknown, but the directions are known. Chandraker and Ramamoorthi [21] used a semi-parametric approach to estimate material reflectance properties from a single image. Our work is related to their approach, but a fundamental difference is that they assume that the reflectance characteristics of the object remain largely stable over θ_d. As we will prove in this chapter this usually is not the case, especially when the lighting direction during sampling is very different from the relighting direction.

Non-parametric approaches Non-parametric representations allow for a greater accuracy and generality. This is also enabled by the availability of comprehensive BRDF databases like MERL [99]. Recent research is shifting towards this direction. Romeiro and Zickler [126] used non-parametric approaches to estimate reflectance under natural illumination, by marginalizing over a distribution of possible lighting environments to cope with the ambiguity between reflectance and illumination. In order to circumvent the color constancy problem, their method only estimates a monochrome reflectance, which leads to limitations when predicting the appearance of objects, such as incorrect colors in highlights. Nöll *et al.* [109] started from a sparsely measured input and used the concept of correction functions to solve for the full BRDF, also handling outliers. The environment lighting in [126] or the sparsely sampled input in [109] already provide many samples of the BRDF, which are - most importantly - scattered across the reflectance space. Although these methods work well for a sparsely sampled BRDF, when the input samples are concentrated in a narrow space of the BRDF domain, as in our case, they tend to overfit the input samples, thereby distorting colors under grazing angles (see Sec. 4.4).

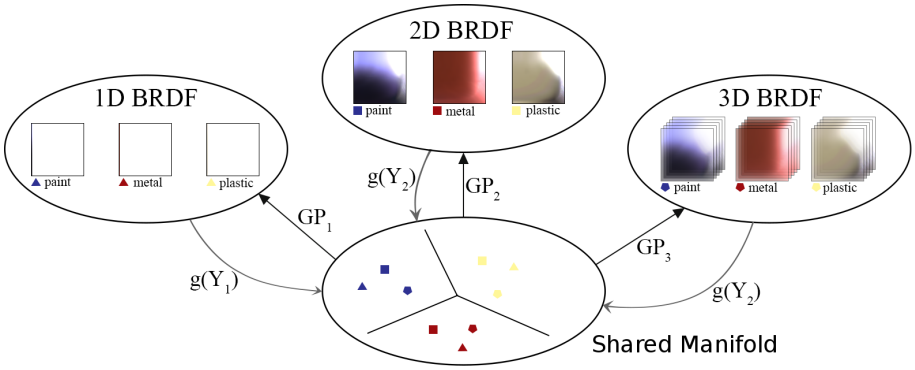


Figure 4.1: Within the DS-GPLVM, BRDFs of different dimensionality (1D, 2D, 3D) can be regressed to a shared manifold. Starting from a single 1D BRDF one can extrapolate to 2D, 3D models.

4.3 Method

4.3.1 Problem Formulation

In this chapter, we consider 1D, 2D, and 3D simplifications of the BRDFs. We aim at inferring higher-dimensional BRDFs from lower-dimensional ones. In particular, from the measured 1D BRDF slice (using a camera with flash), we want to infer the complete 2D or 3D BRDF. This said, we formulate the problem as generally as possible, as the same principles could be used for the transition among differently dimensioned BRDFs as well.

In order to learn how such inference should take place, we use a training set of different materials, for which we can derive their 1D, 2D, 3D, *etc.* BRDFs. We assume to have N such samples (materials). In order to arrive at our general formulation, we assume we have BRDFs from dimension 1 up to V. The entire training set is written as $\mathbf{Y} = \{\mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(V)}\}$, with $\mathbf{Y}^{(v)} = [\mathbf{y}_1^{(v)}, ..., \mathbf{y}_N^{(v)}]^T \in \mathbb{R}^{N \times D}$ with $v = 1, ..., V$ (i.e. v specifies the dimensionality of a BRDF) and D the size of the observation space. For instance, the θ_h axis has been divided into 90 intervals for each RGB channel, thus for our 1D BRDF slice (all values for $\theta_d = 0$) $D = 90 \cdot 3$. Similarly, the θ_d axis was divided into 90 intervals, resulting in a 2D BRDF with $D = 90 \cdot 90 \cdot 3$. The ϕ_d axis was divided into 180 intervals, yielding $D = 180 \cdot 90 \cdot 90 \cdot 3$ for a 3D BRDF. We then seek to find a low-dimensional shared manifold $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N \times q}$, where $q \ll D$ is the size of the manifold that generates all V-dimensional BRDFs simultaneously. Fig. 4.1 summarizes our approach.

4.3.2 Gaussian Process Latent Variable Model

Within the Shared Gaussian Processes (GPs) framework [136, 40], the joint likelihood of $\mathbf{Y}$, given the shared manifold $\mathbf{X}$, can be factorized as follows:

$$p(\mathbf{Y}|\mathbf{X}, \theta_s) = p(\mathbf{Y}^{(1)}|\mathbf{X}, \theta^{(1)}) \times ... \times p(\mathbf{Y}^{(V)}|\mathbf{X}, \theta^{(V)}), \tag{4.1}$$

where the likelihood of the observed BRDF data for dimension v, given the shared manifold, is given by:

$$p(\mathbf{Y}^{(v)}|\mathbf{X}, \theta) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}^{(v)}|^D}} \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{K}^{(v)})^{-1}\mathbf{Y}^{(v)}(\mathbf{Y}^{(v)})^T))). \tag{4.2}$$

Here, $K^{(v)}$ is the kernel matrix, the elements of which are obtained by applying the covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ to each training data pair $(i, j) \in 1, ..., N$. The covariance function is usually chosen as the sum of the Radial Basis Function (RBF) kernel, bias and noise terms, i.e.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-\frac{\theta_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2) + \theta_3 + \frac{\delta_{i,j}}{\theta_4}, \tag{4.3}$$

where $\delta_{i,j}$ is the Kronecker delta function, and $\theta^{(v)} = (\theta_1^{(v)}, \theta_2^{(v)}, \theta_3^{(v)}, \theta_4^{(v)})$ are the kernel parameters [119]. Each v-dimensional BRDF space is generated from the shared manifold via a separate GP, controlled by the parameters stored in $\theta_s = \theta^{(1)}, ..., \theta^{(v)}$. The shared manifold $\mathbf{X}$ is then obtained as the mean of the posterior distribution $p(\mathbf{X}, \theta_s|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \theta_s)p(\mathbf{X})$, where a prior is usually placed over the manifold. This prior allows us to include our knowledge about the BRDF spaces into the learning task.

4.3.3 Discriminative Shared-Space Prior

The choice of the prior will be explained shortly after describing why such a prior is crucial. As shown in [21], clustering the BRDFs into classes of similar material behaviour (e.g. plastics, paints, synthetic and natural fibers) allows us to leverage the unique reflectance properties of each class of materials. Inspired by their approach, we opted for a discriminative prior that encourages the latent positions of the examples of the same class (e.g. plastics) to be close and those of different classes (e.g. plastics and paints) to be far on the shared manifold. To this end, we chose the discriminative shared-space prior [42], which is based on the graph Laplacian matrix. We start by constructing the dimension-specific weight matrices $\mathbf{W}^{(v)}$, by accounting for the data location along with the class.

Specifically, the elements of the weight matrix $\mathbf{W}^{(v)}$ are obtained by applying the RBF kernel to the BRDF data as:

$$\mathbf{W}_{ij}^{(v)} = \begin{cases} \exp(-\frac{\|\mathbf{y}_i^{(v)} - \mathbf{y}_j^{(v)}\|^2}{t^{(v)}}), & \text{if i} \neq \text{j and } c_i = c_j \\ 0, & \text{otherwise,} \end{cases} \tag{4.4}$$

with $\mathbf{y}_i^{(v)}$ the i-th sample in $\mathbf{Y}^{(v)}$, c_i the class label, and $t^{(v)}$ the kernel width which is set to the mean squared distance of the data. The graph Laplacian for dimension v is then $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$, where $\mathbf{D}^{(v)}$ is a diagonal matrix with $\mathbf{D}_{ii}^{(v)} = \sum_j \mathbf{W}_{ij}$. Since the graph Laplacians of different BRDF dimensions have a varying scale, we normalize them as $\mathbf{L}_N^{(v)} = (\mathbf{D}^{(v)})^{-1/2}\mathbf{L}^{(v)}(\mathbf{D}^{(v)})^{-1/2}$. Hence, the joint (regularized) Laplacian can now be defined as:

$$\tilde{\mathbf{L}} = \mathbf{L}_N^{(1)} + ... + \mathbf{L}_N^{(V)} + \xi\mathbf{I} = \sum_v \mathbf{L}_N^{(v)} + \xi\mathbf{I}, \tag{4.5}$$

where $\mathbf{I}$ is the identity matrix, and ξ a parameter which ensures that $\tilde{\mathbf{L}}$ is positive-definite. The chosen discriminative shared-space prior can finally be determined as:

$$p(\mathbf{X}) = \prod_{v=1}^{V} p(\mathbf{X}|\mathbf{Y}^{(v)})^{\frac{1}{V}} = \frac{1}{V \cdot Z_q} \exp\left[-\frac{\beta}{2}\text{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X})\right]. \tag{4.6}$$

In Eq. 4.6, Z_q is a normalization constant and $\beta > 0$ is a scaling parameter. As stated before, this prior aims at maximizing the class separation in the shared manifold learned from BRDF data of all the different dimensions. Using this prior, the negative log-likelihood of the model is given by:

$$L_s(\mathbf{X}) = \sum_v L^{(v)} + \frac{\beta}{2}\text{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X}), \tag{4.7}$$

with $L^{(v)}$ the negative log-likelihood computed by:

$$L^{(v)} = \frac{D}{2}\ln|\mathbf{K}^{(v)}| + \frac{1}{2}\text{tr}[(\mathbf{K}^{(v)})^{-1}\mathbf{Y}^{(v)}(\mathbf{Y}^{(v)})^T] + \frac{ND}{2}\ln 2\pi. \tag{4.8}$$

To learn both the shared manifold $\mathbf{X}$ and the kernel parameters θ_s we minimize the negative log-likelihood in Eq. 4.7, as will be explained below.

4.3.4 Back-Constraints

The model that was described above finds the shared manifold among the different dimensions (i.e. 1D, 2D, 3D) of the input data (i.e. BRDFs). However,

in order to embed new BRDF samples in the shared manifold, we need to learn the back-mappings from the different BRDF spaces to the shared manifold. These back-mappings constrain the learning of the shared manifold by acting as additional regularizers in the model. Specifically, the data that are close in the original BRDF space are constrained to be close on the manifold too, enforcing the topology of the BRDF space to be preserved on the shared manifold. Therefore, we define V sets of constraints that enforce separate back-mappings for each common dimensionality of the BRDFs to the shared manifold. These constraints, referred to as independent back-projections (IBP), were first introduced in [41], and they are given by:

$$\underbrace{\mathbf{X} = g(\mathbf{Y}^{(v)}, \mathbf{A}^{(v)}) = \mathbf{K}_{bc}^{(v)}\mathbf{A}^{(v)}}_{\textbf{IBP from each BRDF dimension } v=1,...,V}, \tag{4.9}$$

where $g(\cdot,\cdot)$ represents the mapping functions learned using kernel regression. The elements of $\mathbf{K}_{bc}^{(v)}$ are calculated by $k_{bc}(\mathbf{y}_i, \mathbf{y}_m) = \exp(-\frac{\gamma}{2}\|\mathbf{y}_i - \mathbf{y}_m\|^2)$ with γ being the inverse width of the kernel. In what follows, we present the algorithm that simultaneously learns the shared space and back-mappings in the model.

4.3.5 Model Learning

To learn the model parameters we minimize the negative log-likelihood in Eq. 4.7 w.r.t. the IBP constraints:

$$\begin{gathered}\min_{\mathbf{X},\theta_s,\mathbf{A}} L_s(\mathbf{X}) + R(g), \\ \text{IBP}(\mathbf{X}, \mathbf{A}^{(v)}) \triangleq \mathbf{X} - \mathbf{K}_{bc}^{(v)}\mathbf{A}^{(v)} = 0, v = 1, ..., V\end{gathered} \tag{4.10}$$

where $R(g)$ is the regularizer defined in the space of $g(\cdot,\cdot)$. The optimal functional form of $R(g)$ can be obtained by applying the Representer Theorem [131], and is given by:

$$R(g) = \sum \frac{\lambda^{(v)}}{2} r(g^{(v)}),\ r(g^{(v)}) = \text{tr}((\mathbf{A}^{(v)})^T \mathbf{K}_{bc}^{(v)} \mathbf{A}^{(v)}). \tag{4.11}$$

4.3.6 Parameter Optimization

In the following paragraphs we present the optimization procedure as originally described in [42]. To find the model parameters we need to iteratively solve a set of sub-problems. This is due to the fact that the back-mapping from each BRDF dimensionality can be written as an independent set of linear constraints

(see Eq. 4.10). We begin by using the Lagrange multipliers to integrate the IBP constraints into the regularized log-likelihood of Eq. 4.10, which in turn results in the Augmented Lagrangian (AL) function:

$$\mathcal{L}^{IBP}(\mathbf{X}, \{\mathbf{A}^{(v)}, \mathbf{\Lambda}^{(v)}\}_{v=1}^{V}) = \\ L_s(\mathbf{X}) + R(g) + \sum_{v=1}^{V} \langle \mathbf{\Lambda}^{(v)}, IBP(\mathbf{X}, \mathbf{A}^{(v)}) \rangle + \frac{\mu}{2} \sum_{v=1}^{V} \| IBP(\mathbf{X}, \mathbf{A}^{(v)}) \|_F^2, \tag{4.12}$$

with $\mathbf{\Lambda}^{(v)}$ the Lagrange multiplier for dimensionality v, $\langle \cdot, \cdot \rangle$ the inner product, and μ a penalty parameter. Since the objective function (see Eq. 4.12) is separable, we can use the Alternating Direction Method (ADM) [14] to decompose it into sub-problems. The use of ADM allows us to alternate between learning the shared manifold and learning the back-mappings for each BRDF dimensionality. Specifically, we first solve for $\mathbf{X}, \theta_s$:

$$\{\mathbf{X}, \theta_s\}_{t+1} = \arg\min_{\mathbf{X}, \theta_s} L_s(\mathbf{X}) + \frac{\mu_t}{2} \sum_{v=1}^{V} \| IBP(\mathbf{X}, \mathbf{A}_t^{(v)}) + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t} \|_F^2, \tag{4.13}$$

we then solve for $\mathbf{A}^{(v)}$ for each dimensionality $v = 1, ..., V$:

$$\mathbf{A}_{t+1}^{(v)} = \arg\min_{\mathbf{A}^{(v)}} r(\mathbf{A}^{(v)}) + \frac{\mu_t}{2} \| IBP(\mathbf{X}_{t+1}, \mathbf{A}^{(v)}) + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t} \|_F^2, \tag{4.14}$$

and finally update the Lagrangian and penalty terms:

$$\mathbf{\Lambda}_{t+1}^{(v)} = \mathbf{\Lambda}_t^{(v)} + \mu_t IBP(\mathbf{X}_{t+1}, \mathbf{A}_{t+1}^{(v)}) \\ \mu_{t+1} = \min(\mu_{max}, \rho \mu_t) \tag{4.15}$$

The problem in Eq. 4.13 lacks a closed-form solution. Therefore, in order to minimize the objective function w.r.t. the shared manifold $\mathbf{X}$ and the kernel parameters θ_s we employ the Conjugate Gradient algorithm (CG) [119]. The problem in Eq. 4.14 resembles the regularized Kernel Ridge Regression (KRR) [141] and its closed-form solution is given by:

$$\mathbf{A}^{(v)} = (\mathbf{K}_{bc}^{(v)} + \frac{\lambda^{(v)}}{\mu_t} \mathbf{I})^{-1} (\mathbf{X} + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t}). \tag{4.16}$$

As this solution is dependent on the parameters $\gamma^{(v)}$ and $\lambda^{(v)}$ solving for it directly would require costly cross-validation procedures. Instead, we can use the Leave-One-Out (LOO) cross-validation procedure for the KRR to learn the parameters $\gamma^{(v)}$ and $\lambda^{(v)}$ and then obtain $\mathbf{A}^{(v)}$ indirectly. The goal of LOO is

to minimize the difference between the prediction $\hat{\mathbf{x}}_i^{(-i)}$ (the superscript here denotes that the i-th sample is left out) and the actual output $\mathbf{x}_i$ for all samples. For this, we first define the matrix

$$\mathbf{M} \triangleq \begin{bmatrix} m_{ii} & \mathbf{m}_i^T \\ \mathbf{m}_i^T & \mathbf{M}^i \end{bmatrix} = (\mathbf{K}_{bc}^{(v)} + \frac{\lambda^{(v)}}{\mu_t}\mathbf{I}) \tag{4.17}$$

where the inverse matrix from Eq. 4.16 is partitioned so that the elements corresponding to the i-th sample appear only in the first row and column of $\mathbf{M}$ ($\mathbf{X}$ and $\mathbf{\Lambda}_t^{(v)}$ are also re-ordered to have the i-th row on top). We also denote with $\mathbf{M}_i = (\mathbf{K}_{bc\setminus i}^{(v)} + \frac{\lambda^{(v)}}{\mu_t}\mathbf{I}_{N-1})$ the kernel matrix formed from the remaining elements. From Eq. 4.16, the prediction and actual target for sample i are:

$$\begin{aligned} \hat{\mathbf{x}}_i^{(-i)} &= \mathbf{m}_i^T \mathbf{M}_i^{-1} \mathbf{m}_i \mathbf{A}_i^{(v)} + \mathbf{m}_i^T \mathbf{A}_{-i}^{(v)} \\ x_i &= m_{ii}\mathbf{A}_i^{(v)} + \mathbf{m}_i^T \mathbf{A}_{-i}^{(v)} - \mathbf{\Lambda}_i^{(v)}/\mu_t \end{aligned} \tag{4.18}$$

and then the cost of the LOO procedure can be defined as:

$$E_{LOO} = \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i^{-i}\|^2 = \frac{1}{2}\sum_{i=1}^{N} \|\frac{\mathbf{A}_i^{(v)}}{[\mathbf{M}^{-1}]_{ii}} - \frac{\mathbf{\Lambda}^{(v)}}{\mu_t}\|^2 \tag{4.19}$$

As a final step, we minimize E_{LOO} with respect to the parameters $\gamma^{(v)}$ and $\lambda^{(v)}$ using CG, and then obtain $\mathbf{A}^{(v)}$ from Eq. 4.16.

4.3.7 Model Inference

To perform inference in the described model, we first project the test data $\mathbf{y}_*^{(v)}$ from a single BRDF dimensionality space $\mathbf{Y}^{(v)}$ (e.g. 1D BRDFs) to the shared manifold using Eq. 4.9. As a result we get the projections $\mathbf{x}_*$ in the latent space. Finally, from the shared manifold we can move back to the other BRDF dimensionality spaces $\mathbf{Y}^{(-v)}$ (i.e. inferring 2D/3D BRDFs from the 1D slices) using the forward-mappings:

$$\begin{aligned} \mathbf{y}_*^{(-v)} &= (\mathbf{K}_{bc}^{(-v)})_*^T (\mathbf{L}^T \backslash (\mathbf{L} \backslash \mathbf{Y}^{(-v)})) \\ \mathbf{L} &= chol(\mathbf{K}_{bc}^{(-v)} + (\sigma_n^{(-v)})^2 \mathbf{I}) \end{aligned} \tag{4.20}$$

with $chol(\cdot)$ being the Cholesky factorization, and σ_n the noise term. Alg. 1 summarizes the learning and inference.

Learning
Inputs: $\mathcal{D} = (\mathbf{Y}^{(v)}, \mathbf{c}), v = 1, ..., V$
Initialize $\mu_{max} \gg \mu_0 > 0, \rho = const., \mathbf{X}_0, \mathbf{A}_0^{(v)}, \mathbf{\Lambda}_0^{(v)}$
repeat
Step 1: Update $(\mathbf{X}, \theta_s)$ by minimizing Eq. 4.13
Step 2: Minimise E_{LOO} from Eq. 4.19 w.r.t. $(\gamma^{(v)}, \lambda^{(v)})_{v=1,...,V}$
Step 3: Update $(\mathbf{\Lambda}^{(v)}, \mu, \mathbf{A}^{(v)})$ from Eq. 4.15- 4.16
until convergence of Eq. 4.12
Outputs: $\mathbf{X}, \mathbf{A}$

Inference
Inputs: $\mathbf{y}_*^{(v)}$
Step 1: Find the projection $\mathbf{x}_*$ from the observation space (v) to the latent space using Eq. 4.9
Step 2: Estimate the forward-mappings from the latent space to the other observation spaces $(-v)$ using Eq. 4.20
Outputs: $\mathbf{y}_*^{(-v)}$

Algorithm 1: Model: Learning and Inference

4.4 Results

In this section we demonstrate how our method performs compared to existing parametric, semi-parametric and non-parametric approaches on various examples, both synthetic and real. Given all samples of a BRDF database (i.e. MERL), we first select the 1D ($f(\theta_h)$) and corresponding 2D ($f(\theta_h, \theta_d)$) and 3D ($f(\theta_h, \theta_d, \phi_d)$) BRDF representations to form three separate BRDF spaces ($\mathbf{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{Y}^{(3)}\}$). All reflectance samples are converted to the logarithmic scale (i.e. we apply the natural logarithm), to make sure that the processing is not biased towards differences in the higher intensity ranges [99]. Each BRDF is consequently transformed to the CIE LAB color space [44], which is perceptually uniform, meaning that a change of the same amount in a color value should produce a change of about the same visual importance. In order to define the class labels for the discriminative shared space prior we clustered the MERL samples into groups of similar statistical behavior, using Spectral Clustering [112], resulting in a set of two clusters that contain 50 materials each, and they happen to represent very well the 'specular' and 'Lambertian' materials. The weight for the prior β was experimentally set to 50. For the initialization of the shared manifold $\mathbf{X}$ we performed Principal Components Analysis (PCA) on the concatenated matrix of all 3 BRDF spaces $\mathbf{Y}$ and kept

the amount of latent variables that explains 95% of the variance in the data. All our experiments were performed using 5-fold cross-validation: we consider 20 samples out of 100 as the testing set. We used a separate validation set of 20 samples to avoid overfitting the training samples. Consequently, for a single experiment, 60 samples (out of the 100 MERL materials) are used for training, 20 for validation and 20 for testing. In total we carried out 25 experiments using a different random set of training, validation and testing samples each time and kept the mean-performing sample with respect to the chosen error metric (i.e. logged data in CIE LAB color space).

For the given problem we evaluated related methods in literature: From the parametric approaches we chose the method of Ashikhmin *et al.* [5] that uses Schlick's model [130] to represent the Fresnel effect. This model uses a fifth order approximation to describe the BRDF's behavior over θ_d, $F(\theta_d) = r_0 + (1 - r_0)(1 - cos\theta_d)^5$. For the following comparison we assume that the 1D BRDF can be perfectly represented, i.e. $\rho(\theta_h)$ introduces zero error and we examine the performance of the model with respect to the prediction over θ_d. In fact for the 1D BRDF one could use any other parametric model, e.g. [15, 78, 107, 156, 25, 130, 59]. For the semi-parametric approach we opted for the method of Chandraker and Ramamoorthi [21]. In contrast to parametric models that assume that both directions (half-angle, back-scatter direction, etc) as well as the form of the distribution (Gaussian, Beckmann, etc) are known in advance, in this semi-parametric approach reflectance is expressed as a sum of (unknown) univariate non-linear (non-parametric) functions, acting on projections of the surface normal on a few (unknown) directions. They further assume though that when relighting the object for another viewing/lighting configuration these non-linear functions are the same, meaning that the reflectance characteristics of the object over θ_d remain largely stable. Finally, from the non-parametric approaches we compare with the recent work of Nöll *et al.* [109], that used the concept of correction functions to solve for the full 3D BRDF, also handling outliers. For a complete evaluation between several non-parametric methods we refer the reader to [109].

Synthetic evaluation on MERL BRDFs To give a representative evaluation on the performance of our algorithm, we compared the different approaches on the 100 MERL samples. In particular we measured the difference between the ground truth BRDF inputs, and the predicted higher-dimensional BRDFs, starting from a single BRDF slice. To mimic the proposed flash-based system for extracting the 1D BRDF, we used the first BRDF slice where $\theta_d = 0$. For the numerical comparison we used different error metrics, linear $\epsilon_{lin}(x) = x$, square root $\epsilon_{root}(x) = \sqrt(x)$ and logarithmic $\epsilon_{log}(x) = ln(1 + x)$, as well as different color spaces, RGB and CIE LAB. As indicated in [109] the choice of

Table 4.1: Mean and median, RGB and CIE LAB error of all evaluated methods on MERL database for different error metrics. To give a visual impression the table cells are colored from best to worst performance using blue, green, yellow and red respectively. Our method performs consistently better across different error metrics and color spaces.

		2D BRDFs												**3D BRDFs**					
		lin				**root**				**log**				**lin**		**root**		**log**	
		Ours	[109]	[21]	[5]	Ours	[109]	[21]	[5]	Ours	[109]	[21]	[5]	Ours	[109]	Ours	[109]	Ours	[109]
RGB	mean	6.36	29.47	6.78	6.83	0.23	8.91	0.44	0.43	0.07	0.18	0.19	0.19	7.28	31.19	0.26	9.98	0.09	0.20
	median	1.43	5.28	3.53	3.52	0.14	0.26	0.46	0.44	0.05	0.10	0.22	0.25	1.90	5.93	0.16	0.34	0.07	0.11
LAB	mean	6.48	29.38	11.20	10.78	3.25	11.59	5.63	5.03	2.57	5.19	5.36	4.81	7.17	31.05	3.22	12.42	2.50	4.84
	median	4.66	14.60	11.54	11.78	2.70	5.54	5.79	5.19	2.17	4.72	5.03	4.95	5.41	16.20	2.71	5.47	2.20	4.46

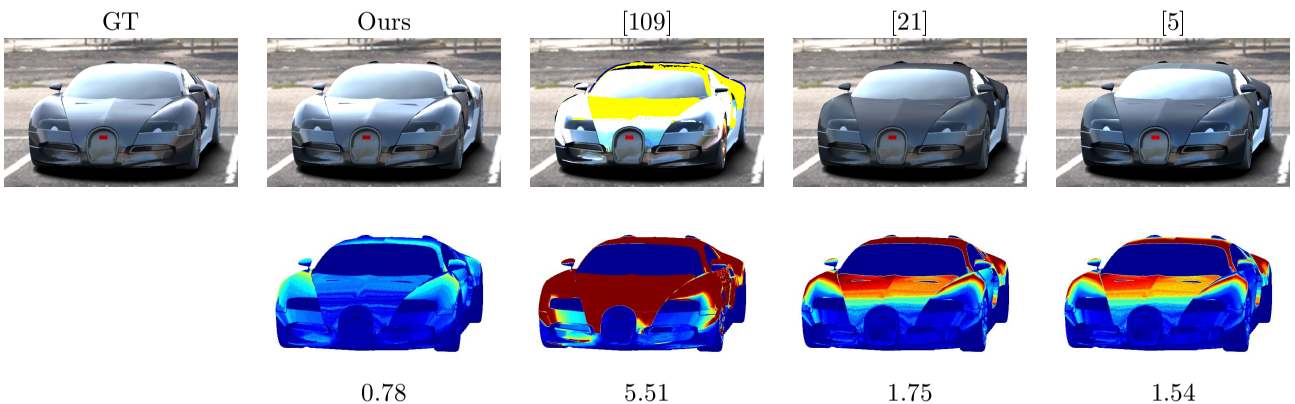


Figure 4.2: Comparison on a synthetic car model with multiple MERL BRDFs rendered under environment lighting. First row: environment renderings, second row: error images in CIE LAB space using the color coded scale of Fig. 4.3, third row: average per pixel error.

the error metric can have a significant impact on the prediction quality. We also considered the mean and median error across the 100 MERL samples. The results are summarized in Tbl. 4.1. For any given error metric or color space, our method outperforms the other approaches.

The general observations are as follows: Assuming that the reflectance characteristics remain stable over θ_d as proposed by Chandraker and Ramamoorthi obviously can not create a proper Fresnel effect. Schlick's approach for the Fresnel approximation used in [5], is only able to partially represent the complicated effects in the grazing angles. The method of Nöll *et al.* generally creates disturbing color artifacts. Of course the latter method was designed to perform well on sparse randomly sampled BRDFs, but for our specific case, it tends to overfit the input 1D slice, resulting in exaggerations. Although our method performs well overall and is consistent with respect to the different error metrics and color spaces, there are cases where the prediction is less successful. Possible failure cases are: (1) material shows Fresnel effects which are not typical for the MERL database, (2) the material shows color changing effects along θ_d, a behavior that can not be deduced from a BRDF slice, (3) the material has a color profile which is not well presented in the MERL database, (4) the test material is not properly clustered (clustering accuracy = 95%). As a final note, simpler linear methods, like [50], could be used for the same task, but our non-linear approach outperforms them and additionally offers a number of advantages, like incorporating BRDFs of different dimensionality in a single manifold and leveraging material information in the learning process.

Synthetic evaluation under environment lighting So far we have measured the differences in the BRDFs themselves. In this section we discuss the effect of the BRDF prediction on environment renders, since surface reflectance properties are clearer and better comparable when objects are viewed under real-world illuminations. In Fig. 4.3 we rendered 2 MERL samples under environment lighting, using the ground truth and predicted BRDFs for all methods. Tbl. 4.2 gives a numerical evaluation, i.e. each output render is compared with the ground truth render; the differences are expressed

Table 4.2: Mean and median CIE LAB error of all evaluated methods on MERL database under environment lighting. To give a visual impression the table cells are colored as in Table 4.1.

		2D BRDFs				**3D BRDFs**	
		Ours	[109]	[21]	[5]	Ours	[109]
pisa field	mean	1.46	3.93	2.89	2.64	1.36	3.52
	median	1.18	1.78	2.63	2.31	1.10	1.84
	mean	1.26	3.68	2.25	2.06	1.17	3.74
	median	0.99	1.41	2.04	1.80	0.98	1.48

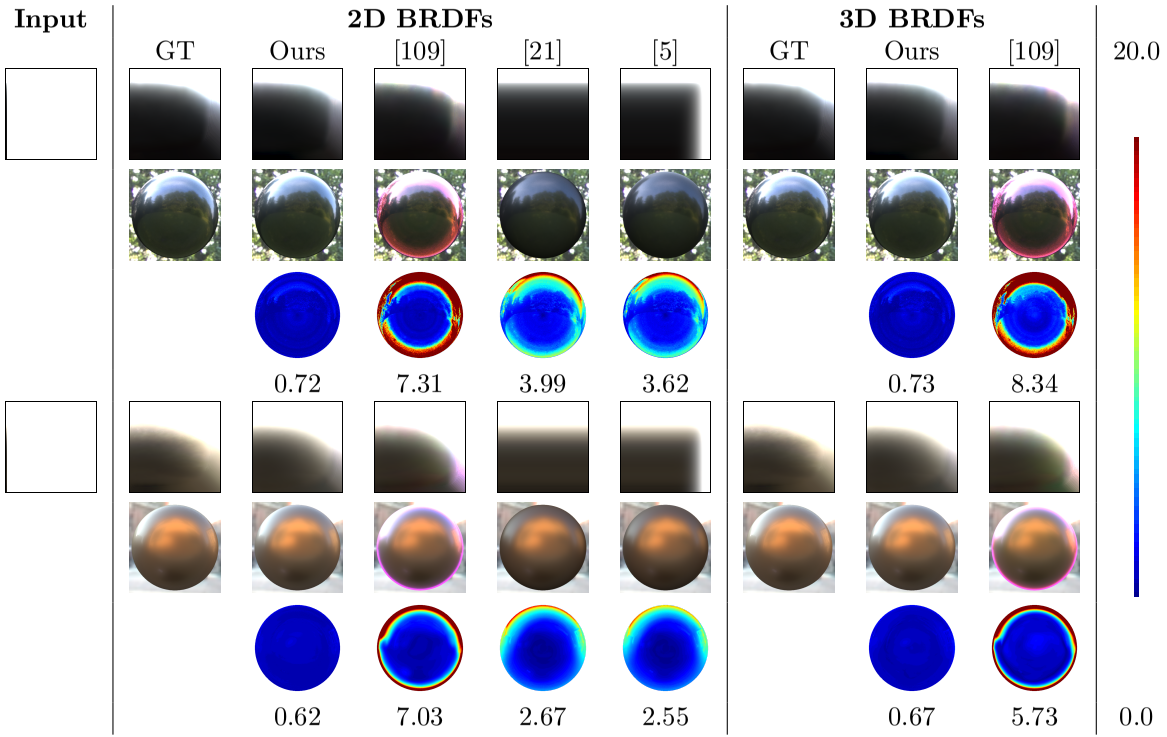


Figure 4.3: Visual and numerical comparison between the predicted BRDFs and their renderings under 2 different environment maps for 2 MERL materials. The first row always shows the BRDFs themselves (for $\phi_d = 90$), the second the environment renderings, the third the error images in CIE LAB space using the color coded scale in the right side of the figure, and the fourth the average per pixel error.

Figure 4.4: BRDF predictions of a real spherical object for the different approaches. Our approach leads to more convincing results in this case too.

in CIE LAB space using logarithmic scale. Mean and median differences over all MERL samples are included in the evaluation. Again, our method outperforms the existing ones both numerically and visually. An overall observation is that the renders are far more natural when a proper Fresnel effect exists in the BRDF prediction, which is generally the case for our method.

Evaluation on real spheres In the next experiment we wanted to evaluate whether our method can be used to measure and render materials in real life circumstances. In particular, we considered a set of reflective spheres. We took HDR pictures using a head-on light. The reflectance samples from this setup provide a single BRDF slice. At the same time, we photographed the same spheres in a real environment, where the environment map itself was scanned separately. Given the 1D BRDF from the initial setup, we predicted the 2D one, rendered it with the scanned environment map and compared with the real-life picture under the same environment. The problem is not straightforward since we had to compensate for effects such as white balance, different color temperatures of the head-on light and the environment, differences between our capturing setup and the one in MERL, but generally we are able to create more convincing results compared to the other methods according to Fig. 4.4.

Application 1: Relighting So far we have considered only spherical objects. In this section we evaluated the method for more realistic applications such as virtual relighting of 3D models. In Fig. 4.2 we consider a car model in a real environment. We selected three metallic MERL samples for the overall body, the hood, and the bumpers. The evaluation carried out is very similar to the one used on the environment renders of the spheres, i.e. we compare the ground truth renders with the predicted BRDF renders for every method. Fig. 4.2 shows the error images in LAB color space. This experiment suggests a possible application where the user samples a material, e.g. from a sphere, using only a head-on light or flash, and through the prediction pipeline one can create a more realistic BRDF representation for photo-realistic rendering.

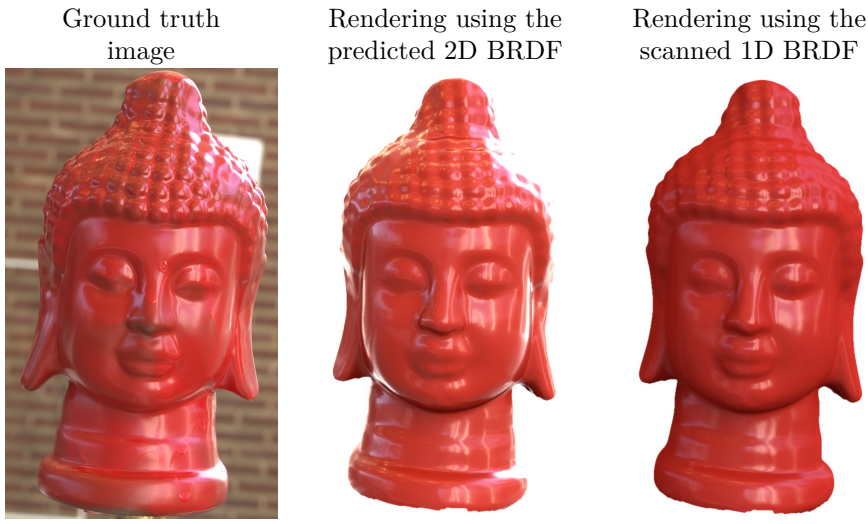


Figure 4.5: A Buddha head model scanned with the flash-based technique of Chapter 3 and rendered at a real-life environment using 1D, 2D BRDF.

Application 2: Flash-based photography As a final experiment, we measured the BRDF of a real object with complex geometry. Fig. 4.5 shows a Buddha statue, the shape of which is extracted using SfM + PS. Given the shape input, the flash-based imagery can be used to derive a BRDF slice. Using our method a 2D BRDF is generated and the model can now be rendered under a real environment. We additionally photographed the Buddha statue in the real environment. Fig. 4.5 shows a visual comparison between the virtual rendered image and the original photograph. This experiment indicates that the assumption of having a known 3D shape is not to be taken too strictly. As shown in Chapter 3 we can efficiently extract 3D shape and BRDF slices just by using a camera with flash.

For extensive results regarding the method presented in this chapter we refer the reader to the supplemental material: `http://homes.esat.kuleuven.be/~sgeorgou/files/ICCV_2015_BRDF_DSGPLVM_Suppl.pdf`.

4.5 Conclusion

We have proposed an approach to infer higher order reflectance information starting from the minimal input of a single BRDF slice. The extensive results

show that our method performs well overall and is consistently better compared to other approaches with respect to the different error metrics and color spaces in both synthetic and real data. This chapter is built upon the previous one and allows to faithfully relight the scanned objects of Chapter 3 for different lighting/viewing configurations than the ones used for scanning, which is the camera/flash combination in this case.

So far, we have explored different approaches for estimating 3D shape and surface reflectance under controlled environmental lighting. In particular, the environmental illumination is only coming from approximately point light sources, like the camera's flash. In the following chapters we allow for less controlled illumination and try to recover surface characteristics and lighting under natural illumination, *e.g.* when the objects of interest are placed in an outdoors scene with light coming from every direction.

Chapter 5

Estimating Environmental Illumination

Recovering the natural illumination from images is a challenging task. Existing approaches in literature rely on strong assumptions to tackle this problem, such as approximating natural illumination with multiple point light sources, assuming an object consists of a single material, or using HDR images as input. On the contrary, we exploit two properties often found in everyday images to remedy this situation without using strong assumptions. First, images rarely show a single material, but rather multiple ones that all reflect the same illumination. In fact, the appearance of each material is observed only for some surface orientations, not all. Second, parts of the illumination are often directly observed in the background, without being affected by reflection. Typically, this directly observed part of the illumination is even smaller.

In this chapter, we address the problem of estimating natural illumination from a single LDR image. In this regard, this chapter addresses Research Question 3: *How can we estimate the environmental illumination of a multi-material object given an image as the sole input?* We propose a deep CNN that combines prior knowledge about the statistics of illumination and reflectance with an input that makes explicit use of the two key observations described above. Our approach maps multiple partial LDR material observations represented as reflectance maps and a background image to a spherical HDR environment map. For training and testing we also propose a new data set comprising of synthetic and real images with multiple materials observed under the same illumination.

The work in this chapter is based on the publication:

- S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars and L. Van Gool, *What Is Around The Camera.* Published in IEEE International Conference on Computer Vision (ICCV) 2017.

5.1 Introduction

Observing a single image, how precisely can we retrieve the omni-directional, incoming illumination under which its foreground objects were photographed (i.e. the environment map)? Intuitively, two partial and imperfect sources of information are available: the light reflected from the visible surfaces and the directly observed background. This line of work is the first to demonstrate how deep learning can be used to combine these cues to resolve a natural estimate of the full illumination. Fig. 1.3 gives a preview.

Traditionally, acquiring the HDR illumination requires placing a mirror ball (light probe) into the scene and capturing images with multiple exposure steps, followed by special post-processing [30]. This is a time-consuming and expensive process known only by experts and is also not an option for already existing footage or dynamic scenes. In this chapter, we drastically reduce the acquisition effort by taking a single LDR photo. Deep learning allows our method to use everyday objects - i.e. far-from-perfect-mirrors both in terms of shape and materials - to act as light probes (cf. the Dino in Fig. 1.3).

This is a challenging task due to the many factors affecting how impinging illumination is turned into object appearance. First, the albedo is unknown, and thus surfaces might *e.g.* appear green because the illumination is green or the albedo is. Next, there is more to reflection than a scalar albedo: light coming from multiple directions may be reflected to different degrees in the direction of the camera, thus further increasing the ambiguity. Finally, the illumination information needs be retrieved in HDR to be of practical use, even if the typical sensor only takes LDR images.

In order to computationally solve this challenge, we exploit the two pieces of information most readily available to us. First, we use the way in which the different materials covering the foreground objects reflect the illumination from the environment. Second, behind the foreground objects we typically observe part of the environment directly as the image's background. These two sources of information tend to provide complementary information about the environment, as it is mainly the part not visible as image background that is reflected by the objects.

As to the reflection by the foreground objects, they rarely are made of a single

material. For instance, the example in Fig. 1.3 shows three materials. Each material reflects the same illumination with a different and unknown BRDF. In practice only a subset of all surface normal orientations are visible for each material, and the estimates of these orientations are noisy. We assume the mapping between surface orientations and appearance (*i.e.* a reflectance map in the sense of [67]) to be known. This can be achieved, either by aligning an existing 3D model to the image, by the use of depth sensors, by extracting per-pixel normals using CNNs [37, 153, 86] or directly, also by means of deep learning [122]. We have designed our system to work with reflectance maps - not an image directly - as input, as to be able to work with all the aforementioned acquisition modalities.

As second piece of information, we exploit parts of the illumination that are often directly visible in the background. While the background is not convolved with a BRDF, it is only a fraction of the full sphere for typical fields-of-view and it is often subject to depth-of-field blur.

We train a deep CNN that combines prior knowledge about the statistics of illumination and reflectance. We also propose a new data set of synthetic and real images consisting of multiple materials under the same illumination. Our CNN observes the LDR appearance of multiple materials represented as reflectance maps, as well as a background image, to produce a full-sphere HDR environment map.

5.2 Previous Work

Object appearance is the result of an intriguing jigsaw puzzle of unknown illumination, material reflectance, and shape. Decomposing it back into these intrinsic properties is far from trivial [10]. Typically, one or two of the intrinsic properties are assumed to be known and the remaining one is estimated. In this chapter, we focus on splitting materials and illumination when the partial reflectance maps of multiple materials seen under the same illumination plus a background image are known. Such an input is very typical in most images, yet not so often studied in the literature.

Key to this decomposition into intrinsic properties is to have a good understanding of their natural statistics. Databases of material reflectance [27, 99, 13] and environmental illumination [29, 35] allow the community to make some first attempts. Yet, exploiting them in practical de-compositions remains challenging.

Reflectance maps *Reflectance maps* [67] assigned appearance to a surface orientation for a given scene, thus combining surface reflectance and illumination. Reflectance maps can be extracted from image collections [56], from a known class [123], or using a CNN [122]. In computer graphics, reflectance maps are used to transfer and manipulate appearance of photo-realistic or artistic "lit spheres" [137] or "MatCaps" [138]. Khan [73] made diffuse objects in a photo appear specular or transparent using image manipulations of the image background that require manual intervention. On the contrary, our approach does not require any user intervention.

Factoring illumination Classic intrinsic images factor an image into shading and reflectance [10]. Larger-scale acquisition of reflectance [99] and illumination [29] have allowed to compute their statistics [35] helping to better solve inverse and synthesis problems. Nevertheless, intrinsic images typically assume diffuse reflectance. Surprisingly, humans do best in recognition of material, shape and illumination on complex geometry, not on plain spheres [150]. As will be shown in Sec. 5.6, our approach indeed shows the same behavior when presented heterogeneous input, which we explicitly target in this work.

Recently, separating material reflectance (henceforth simply referred to as 'material') and illumination was addressed by Johnson and Adelson [69] as well as Lombardi and Nishino [93]. They present different optimization approaches that allow for high-quality estimation of one component if at least one other component is known and remains the same across the image. Instead, we rely on less strict assumptions; the object is made of multiple materials, it can be segmented into its different materials as well as from the background, and the reflectance maps of all materials can be extracted.

Barron and Malik [9] decompose shaded images into shape, reflectance and illumination, but only for scalar reflectance, *i.e.* diffuse albedo, and for limited illumination frequencies. Recently Richter and Roth [125] first estimate a diffuse reflectance map represented in Spherical Harmonics (SH) using approximate normals and then refine the normal map using the reflectance map as a guide. SH are only suitable to represent low-frequency illumination, while our environment maps reproduce fine details.

We address a problem more general than the one of Lombardi and Nishino [93]: they consider a sphere with a single, unknown material on the surface (homogeneous surface reflectance) observed under some unknown natural illumination. As noted in [83, 171, 91] multiple materials help to estimate materials under a single point light source. In this chapter, we ask how multiple materials, instead of a single one, under the same non-point light illumination can help a deep architecture to reason about the lighting. We also work on

partial observations, as in most real applications it is not likely to observe all normals for all materials, but only partial reflectance maps derived from a subset of all normals.

Lombardi and Nishino [92] have used HDR RGB-D images to acquire shape, reflectance and illumination using an optimization-based framework that includes illumination statistics as a prior. We show how HDR illumination can be directly estimated from LDR images of scenes with multiple materials, using deep learning. Barron and Malik [8] made use of similar data to resolve spatially-varying, local illumination. While ours is spatially invariant (distant), we can extract it both with more details, in HDR and from non-diffuse surfaces. In general, previous works have considered HDR input [93, 92], which implies the capture of multiple exposures per image making the capturing process rather impractical, or produced only parametric environment maps [126].

Earlier work has also made use of cues that we did not consider, as they may only be available in some scenes, such as shadows [129]. Lalonde *et al.* [79] have shown how to fit a parametric sky model to a 2D image, but cannot reproduce details such as buildings and trees and exclude non-sky, *i.e.* indoor settings. Karsch *et al.* [71] automatically inferred environment maps by selecting a mix of Nearest Neighbors (NN) from a database of environment maps that can best explain the image assuming diffuse reflectance and normals have been estimated. They demonstrate diffuse relighting but specular materials, that reveal details of a reflection, hardly agree with the input image as seen in our results section. As their data set of illuminations is not publicly available, we have compared to a NN approach based on our own data set of such maps.

Deep learning CNNs have been used for depth [38, 86, 89] and normal estimation [37, 153], as well as intrinsic image decomposition [102, 168] and diffuse illumination estimation [103]. In contrast, we do not estimate geometry, but seek to find detailed non-point light illumination. In addition, our data set contains the combination of HDR environment maps, specular materials and images, which is not well represented in prior recordings (*e.g.* typically assuming diffuse surfaces [102, 168]).

As reflection has similarities to a convolution of illumination and BRDF [118], we also note that deep learning is successful in typical de-convolution tasks, such as super-resolution [32] and removing camera [161] or motion blur [140]. Differently, our de-convolution operates in the spherical illumination domain, with statistics different from images [35] and a kernel typically not found in images: the BRDF.

5.3 Overview

We formulate our problem as learning a mapping from n_{mat} partial reflectance maps [67] and a background image to a single consensual HDR environment map. In particular, we never extract illumination directly from images, but indirectly from reflectance maps. We assume the reflectance maps were extracted using previous work [37, 153, 86, 122]. In our datasets we rely on manually aligned and selected geometry to analyze the limits of what reflectance map decomposition can do and we do not consider the error introduced by the estimated reflectance map itself.

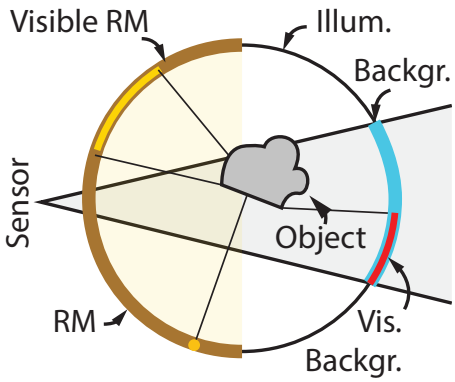


Figure 5.1: Illustration of Sec. 5.3

A reflectance map $L_{\text{o}}(\omega)$ represents the appearance of an object of a homogeneous material under a specific illumination. Under the assumptions of (i) a distant viewer, (ii) distant illumination, (iii) in the absence of inter-reflections or shadows (convex object) and (iv) a homogeneous material, the appearance depends only on the surface orientation ω in camera space and can be approximated as a convolution of illumination and BRDF [118].

The full set of orientations in $\mathbb{R}^3$ is called the 3D *Gauss* sphere Ω (the full circle in Fig. 5.1). Note that, only at most half of the orientations in $\mathbb{R}^3$ are visible in camera space, *i.e.* the ones facing into the direction of the camera. This defines the positive Gauss sphere Ω^+ (the brown half-circle in Fig. 5.1). Also note that, due to the laws of reflections, surfaces oriented towards the viewer also expose illumination coming from behind the camera. The ideal case is a one-material spherical object, that completely contains all observable normals. When its surface behaves like a perfect mirror, that is even better. Then a direct (but partial) environment map is directly observable. In practice, we only observe some orientations for some materials and other orientations for other materials. Sometimes, multiple materials are observed for one orientation, but it also happens that for some orientations, no material might be observed at all. Moreover, the materials tend to come with a substantially diffuse component in their reflectance, thus smearing out information about the environment map. In Fig. 5.1, the brown part shows the half-sphere of the reflectance map and the yellow part within shows the object normals actually observed in the image,

for the example object in the figure.

A second piece of input comes from the background. The visible part of the background in the image shows another part of the illumination, this time from the negative half sphere. In Fig. 5.1, the visible part of the image background is shown in red, the rest - occluded by the foreground - in blue.

The illumination $L_{\mathrm{i}}(\omega)$ we will infer from both these inputs covers the full sphere of orientations Ω (the full circle in Fig. 5.1). Other than the reflectance map, it typically is defined in world space as it does not change when the viewer's pose changes. For the actual computations, both the input (partial reflectance maps and partial background) and the output (illumination) are represented as two-dimensional images using the latitude-longitude parameterization.

The mapping $f := L_{\mathrm{o}} \rightarrow L_{\mathrm{i}}$ we seek to find is represented using a deep CNN. We propose a network that combines multiple convolutional stages - one for each reflectance map, that share weights, and another one for the background - with a joint de-convolutional stage that consolidates the information into a detailed estimate of the illumination.

The training data consists of tuples of reflectance maps l_{o} with a single background image that together form the domain and a corresponding illumination l_{i} that is the range of the mapping learned. We have synthesized a large number of reflection maps of random objects under a random view, with a random material reflectance and random illumination.

We next describe our new dataset in Sec. 5.4 before proceeding to show how it is used for training in Sec. 5.5.

5.4 Dataset

Our dataset consists of synthetic training and testing data (Sec. 5.4.1) and a manually-acquired set of test images of real objects captured under real illumination (Sec. 5.4.2).

5.4.1 Synthetic Data

We now explain how to synthesize train and test data.

Rendering Images are rendered at a resolution of 512×512 using variations of geometry, illumination, materials, and views. The geometry is a random

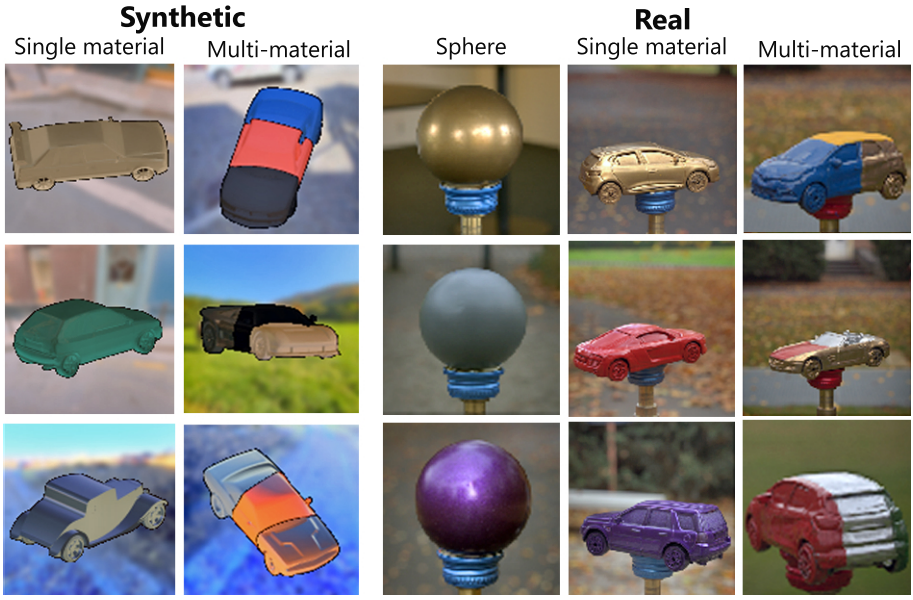


Figure 5.2: Example images from our dataset. **1st col:** Synthetic images of cars with a single material. **2nd col:** Synthetic images of cars with multiple materials. **3rd col:** Photographs of spheres with a single material. **4th col:** Photographs of toy cars with a single material. **5th col:** Photographs of toy cars with multiple materials.

object from the ShapeNet [22] class "car". Later, we show results of our pipeline for both cars and on other shapes though (*e.g.* Fig. 1.3). As large 3D shape datasets from the Internet do not come with a consistent segmentation into materials, we perform a simple image segmentation after rasterization. To this end, we perform k-means clustering ($k = n_{\text{mat}}$) based on positions and normals, both weighted equally and scaled to the range $(-1, 1)$, to divide the shapes into three regions, to be covered with three different 'materials'. Per-pixel colors are computed using direct (no global illumination and shadows) image-based illumination [30]. We also store per-pixel ground-truth positions and normals. As materials we used the 100 BRDF samples from MERL database [99]. The illumination is randomly selected from a set of 105 publicly available HDR environment maps that we have collected. The views are sampled randomly over the sphere, with a fixed field-of-view of 30 degrees. Synthetic examples can be seen at the first two columns of Fig. 5.2.

Extracting reflectance maps The pixel j in the reflectance map of material i is produced by averaging all pixels with material i and orientation ω_j. The final reflectance maps contain 128×128 pixels. These are typically partial with sometimes as little as 10% of all normals observed.

Background extraction The background is easily identified for these synthetic cases, by detecting all pixels where the geometry did not project to. To make the network aware of depth-of-field found in practice, the masked background is filtered with a 2D Gaussian smoothing kernel ($\sigma = 2$).

Building tuples To test our approach with material tuples of arbitrary size n_{mat} while rendering and capturing images, we simply combine n_{mat} random reflectance maps extracted from images with a single material.

Splitting For the single-material case, from the 60 k synthetic images generated, 54 k are used for training and 6 k for testing. Note that, no environment map is shared between the two sets - 94 for training and 11 for testing randomly generated once. For the multi-material case, we used the same protocol as before (identical sets) but this time instead of rendering different car models under the same illumination we render a different part of the same car model (Fig. 5.2).

5.4.2 Real Data

While training can be done on massive synthetic data, the network ultimately is to be tested on real images. To this end, we acquired photographs of both single-material as well as multi-material objects with known geometry under natural illumination which we also captured in HDR (reference).

All images in this set - 112 in total - were used for testing and never for training. Moreover, all 3D models, materials and illuminations in this set are unknown to the train set.

Capture The images are recorded with a common DSLR LDR sensor at a resolution of 20 M pixels and consequently re-scaled to match the training data. For each image, we acquired the environment map too using an HDR image of a spherical mirror. Three variants were acquired: spheres, single-material objects and multi-material objects. For the single-material case, 84 images were taken, showing 6 spheres and 6 toy cars with different materials each and placed under 7 different illuminations. The multi-material data comprises of 30

images, showing 6 different objects (4 cars and 2 non-cars), each painted with 3 materials, captured under 9 different illuminations (6 and 3 respectively). Some materials repeat, as overall 12 different materials were used.

Extracting reflectance maps and background From all images, reflectance maps are extracted in the same way as for the synthetic images. Per-pixel normals are produced using virtual replica geometry from online repositories or scanned using a structured-light scanner. These models were manually aligned to the 2D images. Material and background segmentation was also done manually for all images.

5.5 Network Architecture

Our network consists of three parts (Fig. 5.3) - some of them identical in structure and some sharing weights. First, there is a convolutional *background* network. Second, n_{mat} convolutional *de-reflection* networks that share parameters but run on the reflectance maps of different materials. Third, a final de-convolutional *fusion* network takes as input intermediate stages as well as end results from all reflectance nets, together with the result of the background net, to produce the HDR environment map as an output. All parts are trained jointly end-to-end using an L1 loss on the illumination samples, after applying the natural logarithm and converting them to CIE LAB space. We have experimentally found that these choices nicely balance between learning the dynamic range and the color distribution of the environment map.

Background network Input to the background network (blue part in Fig. 5.3a) is a LDR background image in full resolution *i.e.* 128×128 converted to CIE LAB space. The output is a single, spatially coarse encoding of resolution 4×4. The reduction in spatial resolution is performed as detailed in Fig. 5.4, left. Only the final output of the encoding step will later contribute to the fusion (Fig. 5.3d).

De-reflection network The de-reflection network (blue parts in Fig. 5.3b) consumes partial, LDR reflectance maps also converted to CIE LAB space, where undefined pixels are set to black. It has the same structure as the background network. It starts with the full, initial reflectance map at a resolution of 128×128 and reduces to a spatial resolution of 4×4. We can support an arbitrary, but known and fixed number of materials n_{mat}, as the network needs to be trained for a specific number. In any case, the de-reflection networks are trained with shared parameters (siamese architecture; locks in Fig. 5.3). We want each of

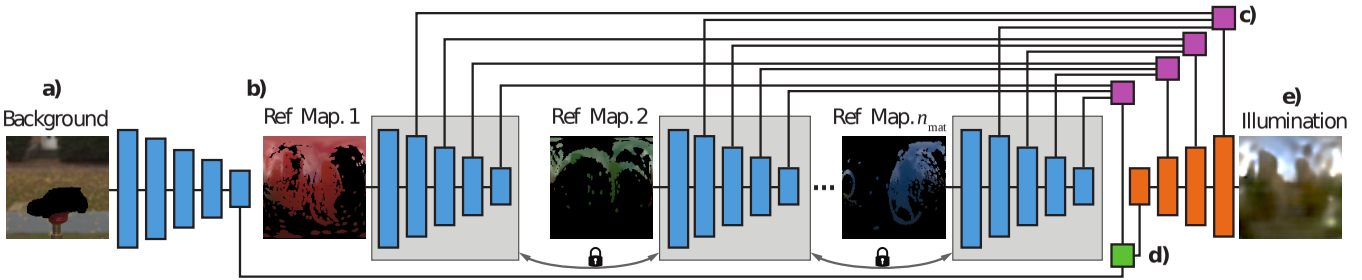


Figure 5.3: CNN architecture of our approach *(from left to right)*. First, the background image is encoded using one independent sub-network *(blue)*. Next, each partial reflectance map is encoded using n_{mat} de-reflection sub-networks that share parameters *(blue)*. Finally, these two sources of information are fused in a de-convolution network *(orange)*. Here, information from all levels of the partial reflectance maps is included *(violet)* as well as the global encoding of the background *(green)*. Details of each sub-network are discussed in the text of Sec. 5.5 and in Fig. 5.4.

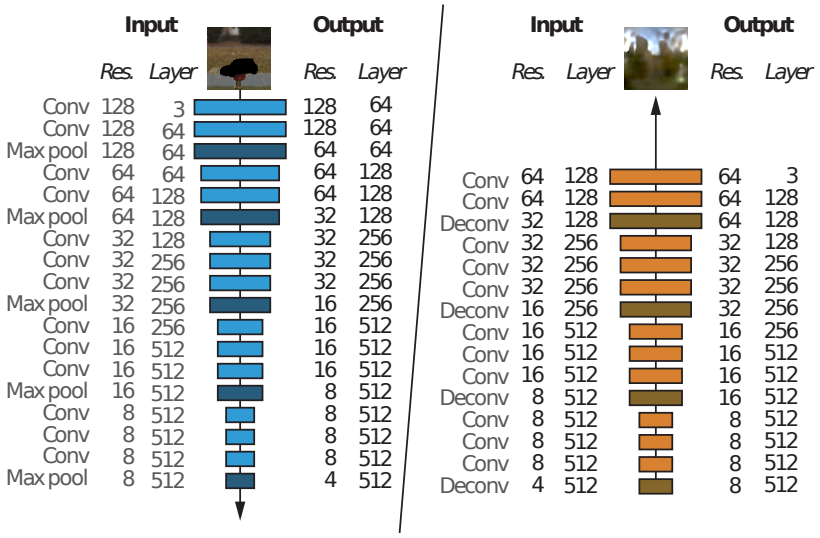


Figure 5.4: Details of the blue and orange sub-networks from Fig. 5.3.

these networks to perform the same operations and do not come in a particular order.

Fusion network The fusion network (Fig. 5.3e) combines the information from the background and the de-reflection networks. The first source of information are the intermediate representations from the reflectance maps (violet, Fig. 5.3c). They are combined using plain averaging with equal weights. This is done at each scale of the de-reflection, respectively, at each level of the fusion. The second source of information is the background (green in Fig. 5.3d). Here, only a single, spatial level is considered, i.e. that of its output. This encoding is concatenated with the average of the encodings from all reflectance maps on the coarsest level (*i.e.* their spatial resolution matches). Result of this sub-network is the final 64×64 HDR environment map (Fig. 5.4).

The receptive field of consecutive convolutional or de-convolutional filters is 3×3 pixels whereas for max pooling filters it is 2×2 pixels. We train this network end-to-end for 100 epochs using MatConvNet with $n_{\text{mat}} = 3$.

5.6 Results

In this section we present both quantitative results (Sec. 5.6.1) that compare different variants or alternative approaches in terms of numbers as well as qualitative results (Sec. 5.6.2) showing possible applications.

5.6.1 Quantitative Evaluation

We quantify to which extent our approach can acquire HDR illumination from LDR photos. As evaluation metric we use the perceptualized DSSIM [154] (less is better). This metric captures the structural similarity between images [23, 113, 102, 122], that is of particular importance when the environment's reflection is visible in a specular surface, such as the ones we target in this chapter.

Model variants and baselines The results of different variants of our approach and baseline methods are presented in terms of performance (Tbl. 5.1) and visual quality (Fig. 5.5):
• Singlet uses only a single reflectance map, *i.e.* our de-reflection network with $n_{\text{mat}} = 1$, but without background.
• Singlet+BG also uses a single reflectance map, as before, but includes the background network too.
• Best-Of-Singlets executes the $n_{\text{mat}} = 1$ de-reflection-plus-background network for each singlet of a triplet individually and then chooses the result closest to the reference by an oracle (we mark all oracle methods in gray).
• Nearest Neighbor picks the nearest neighbor to ground-truth from the training data by an oracle so that the error is minimized. This is an upper bound on what any approach that can only retrieve environment maps from the training data can achieve, like [71].
• Mask-Aware Mean executes $n_{\text{mat}} = 1$ de-reflection-plus-background network for each singlet of a triplet individually and then averages the predicted environment maps based on the sparsity masks of the input reflectance maps.
• Triplet combines three reflectance maps via our de-reflection network with $n_{\text{mat}} = 3$, without background.
• Triplet+BG represents our full model that combines the de-reflection (with $n_{\text{mat}} = 3$) and background network.

Quantitative results All variants are run on all subsets of our test set: synthetic and real, both single and multi-material, for all objects. Results are summarized in Tbl. 5.1. For the synthetic cars, we see a consistent improvement

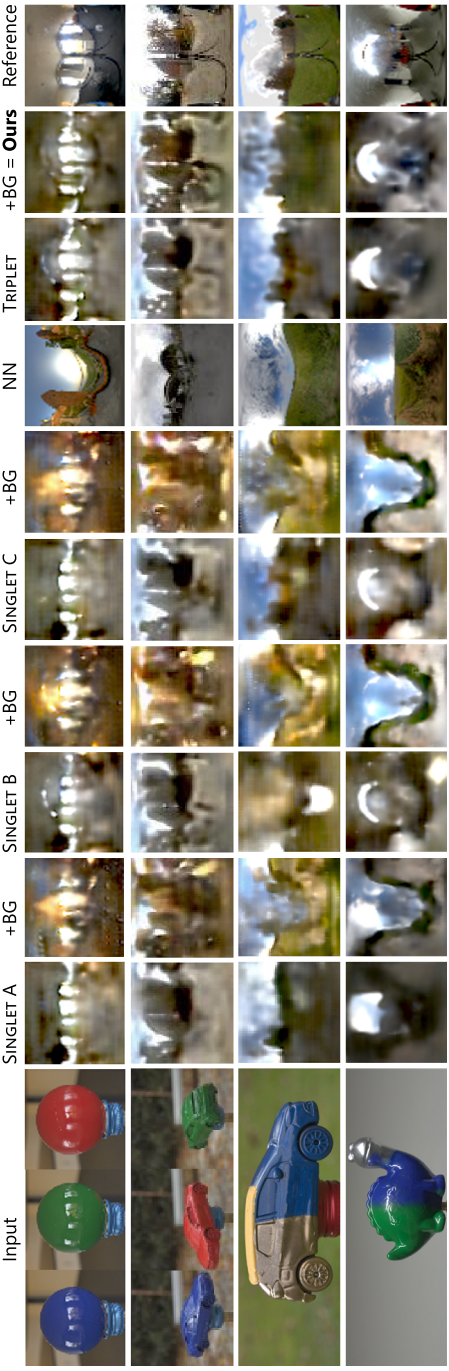
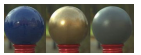


Figure 5.5: Alternative approaches *(left to right)*: **1)**: The input. **2, 4 and 6):** Our approach for $n_{\text{mat}} = 1$. **3, 5 and 7):** the same, including a background. **8):** the nearest neighbor to the reference in the set of all training environment maps. **9):** Our approach for $n_{\text{mat}} = 3$. **10):** Our approach for $n_{\text{mat}} = 3$ and a background, *i.e.* the full approach. **11):** reference. For a quantitative version of this figure see Tbl. 5.1.

Table 5.1: DSSIM error (less is better) for different variants *(rows)* when applied to different subsets of our test set *(columns)*. The best alternative is shown in **bold**. Oracle analysis using ground-truth information are shown in gray. Variant images are seen in Fig. 5.5.

	Synthetic		Real			
						
	Cars (Single)	Cars (Multi)	Spheres	Cars (Single)	Cars (Multi)	Non-cars
Singlet	.311±.011	.316±.011	.324±.002	.337±.002	.335±.005	.315±.002
Singlet + back.	.281±.010	.277±.008	.360±.003	.360±.002	.366±.005	.341±.002
Best-of-singlets	.304±.011	.307±.011	.314±.001	.330±.002	.324±.004	.312±.004
Near. neigh.	.277±.009	.277±.009	.360±.002	.360±.002	.332±.007	.313±.004
Mask-aware mean	.290±.012	.293±.012	.306±.002	.324±.002	.305±.004	.285±.002
Triplets	.268±.011	.277±.011	.313±.001	.332±.002	.284±.002	.288±.001
Triplets + back.	**.210±.007**	**.226±.007**	**.305±.001**	**.315±.001**	**.272±.004**	**.279±.001**

by adding background information already for the SINGLET - even outperforming BEST-OF-SINGLETS. Across all experiments, there is consistent improvement from SINGLET to TRIPLET to TRIPLET+BG. TRIPLET+BG has consistently the best results - in particular outperforming the **Nearest Neighbor**, which indicates generalization beyond the training set environment maps as well as the hand-crafted fusion scheme MASK-AWARE MEAN. Overall, it is striking that performance for the multi-material case is very strong. This is appealing as it is closer to real scenarios. But it might also be counter-intuitive, as it seems to be the more challenging scenario involving multiple unknown materials with less observed orientations. In order to analyze this, we first observe that for SINGLET, moving from the single to the multi-material scenario does not affect performance much. We conclude that our method is robust to such sparser observation of normals. More interestingly, our best performance in multi-material scenario is only partially explained by exploiting the "easiest" material, which we see from BEST-OF-SINGLETS. The remaining margin to TRIPLET indicates that our model indeed exploits all 3 observations and that they contain complementary information.

Visual comparison Example outcomes of these experiments, are qualitatively shown in Fig. 5.5. For tone-mapping, the .90-percentile is used to find a reference exposure value. We then apply the same tone-mapper with this authoritative exposure to all alternatives, including ours. Horizontally, we see that individual reflectance maps can indeed estimate illumination, but contradicting each other and somewhat far from the reference (columns labeled SINGLET in Fig. 5.5). Adding the BG information can improve color sometimes (columns +BG in Fig. 5.5). We also see that a nearest neighbor approach (column NN in Fig. 5.5) does not perform well, even if it was feasible. Proceeding with triplets (column TRIPLET in Fig. 5.5) gets closer to the true solution, but only adding the background (OUR in Fig. 5.5) results in the best prediction. We see that as the difficulty increases from spheres over single- and multi-material to complex shapes, the quality decreases while a plausible illumination is produced in all cases. Most importantly, the illumination can also be predicted from complex, non-car multi-material objects such as the dinosaur or pig geometry as seen in the last column. We refer the reader to the supplementary material[1]for a complete visualization of all alternatives across the whole test data set.

Varying the number of materials In another line of experiments we look into variation of n_{mat} in Tbl. 5.2. Here the number of input reflectance maps increases from 1 up to 5. In each case we include the background and run both on spheres and single-material cars, for which these data are available for $n_{\text{mat}} > 3$. Specifically, we use the real singlets, that we combine into tuples of reflectance maps according to the protocol defined in Sec. 5.4. We see, that

Table 5.2: Reconstruction error on different number of materials n_{mat}.

	Spheres	Cars (Single)
Singlet + Back.	.360±.003	.360±.002
Doublets + Back.	.320±.002	.327±.002
Triplets + Back.	.305±.001	.315±.001
Quadruplets + Back.	.309±.001	.306±.001
Quintuplets + Back.	**.292±.001**	**.295±.001**

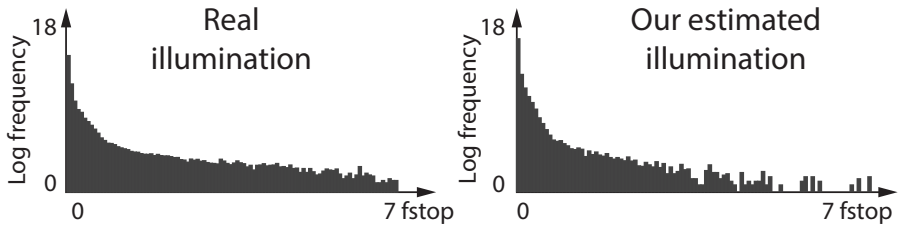


Figure 5.6: Histogram of log luminance *(vertical)* plotted over bins of f-stops *(horizontal)*. An LDR image spans roughly 2-3 f stops.

although we have not re-trained our network but rather copy the shared weights that were learned using $n_{\text{mat}} = 3$ materials, our architecture does not only retain efficiency across an increasing number of materials in both cases, but in fact uses the mutual information to produce even an increase in quality. This is in agreement with observations that humans are better in factoring illumination, shape and reflectance from complex aggregates than for simple ones [150].

Analyzing predicted dynamic range Finally, we have plotted the distribution of luminance over the test data set and compare it to the distribution of the illuminations we estimate in Fig. 5.6. We see, that our approach reproduces the full-dynamic range of luminances although it operates using only LDR inputs. In the higher range however, we do not reproduce some brighter values found in the reference. This indicates, that our results are both favorable in structure as seen from Tbl. 5.1 and Tbl. 5.2 as well as according to more traditional measures such as log L1 or L2 norms.

5.6.2 Qualitative Evaluation

The visual quality is best assessed from Fig. 5.7, that shows, from left to right, the complete input information (a, b), the intermediate stages (c), our result (d)

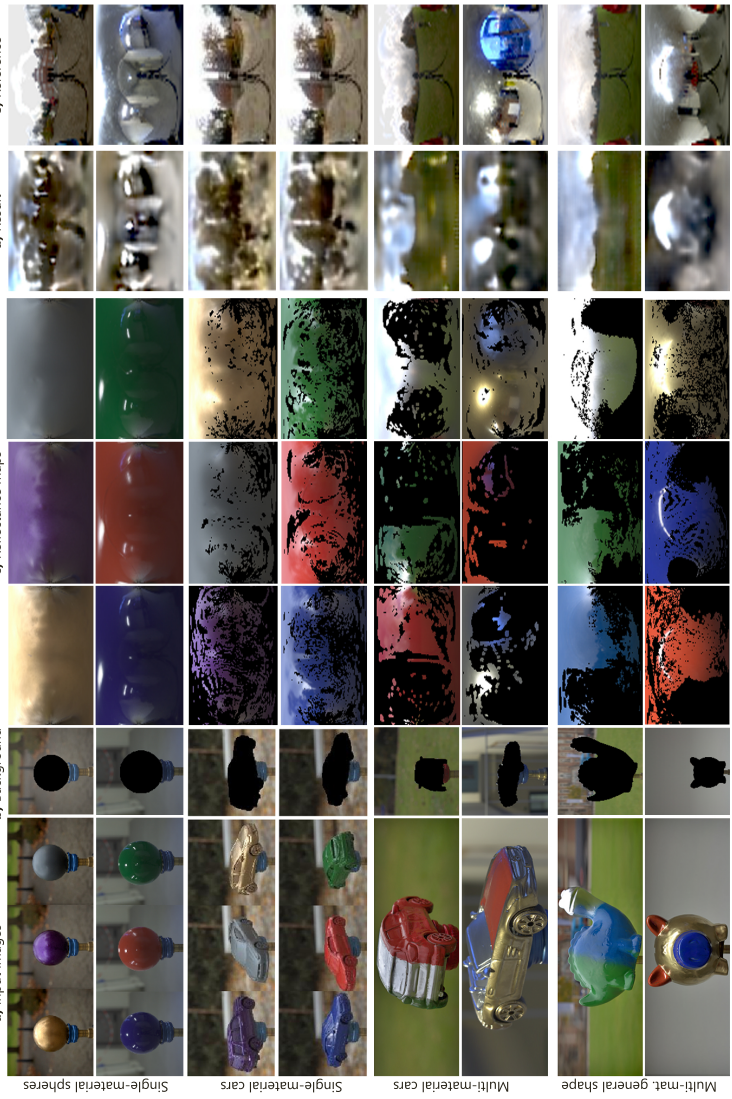


Figure 5.7: Results on real objects (rows): spheres, single-material cars, multi-material cars and non-cars. *a:* input LDR images. *b:* extracted background. *c:* estimated reflectance maps. *d:* predicted illumination. *e:* the ground truth.

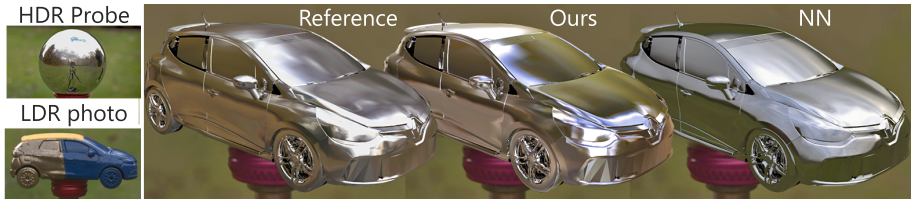


Figure 5.8: Comparison of re-rendering using the reference, ours, and nearest neighbor for a specular material. Ours is more similar to the reference, while not requiring to acquire an HDR light probe.

and the ground-truth environment map as reference (e). The difficulty increases vertically: Starting from spheres, we proceed to scenes that combine three single material objects over single objects with multiple materials to non-car shapes with multiple materials. This shows how non-car shapes at test time can predict illumination, despite training was done on cars and car parts. We see how the reflectance map information is partial and contradicting, but still it can be disambiguated and consolidated into a reasonable estimate of illumination as seen from comparing the two last columns.

To get an idea not only about the improvement but also about the effectiveness in a real application, we show how inserting a virtual object with a new material looks like when illumination is captured using our approach vs. a light probe (Fig. 5.8). In the traditional setup, as used in acquiring test data we encounter multiple exposures, (semi-automatic) image alignment, a mirror ball with known reflectance and geometry. In our approach we have an unknown object with unknown material and a single LDR image. Note how similar image and rendered results are. This is only possible when the HDR is correctly acquired. At the same time, a nearest-neighbor oracle approach, that is a bound above anything achievable in practice already performs worse: The reflection alone is plausible, but far from the reference. Please see the supplemental video[1] for more such applications.

5.7 Conclusion

We have shown an approach to estimate natural illumination in HDR when observing a shape with multiple, unknown materials captured using an LDR sensor. We phrase the problem as a mapping from reflectance maps to environment maps that can be learned by a suitable novel deep convolution-

[1]Link: `http://homes.esat.kuleuven.be/~sgeorgou/multinatillum/index.html`

de-convolution architecture we propose. Training and evaluation are both made feasible thanks to a new data set combining both synthetic and acquired information.

Despite of the ability of the presented method to estimate a HDR environment map from a single LDR image, some fundamental questions arise: Do we really need to know the geometry of the object in order to estimate its reflectance map(s)? What if the latter is unknown? For example, a 3D model of the object is not found in repositories or an RGB-D sensor is not available. In this case, to what extend can we retrieve the surface reflectance and the environmental illumination and under which assumptions?

The following chapter directly addresses these questions and presents a method for estimating parametric reflectance and natural illumination information from a single LDR image. In contrast to Chapters 3, 4 and 5, however, we do not assume one or more components (shape, reflectance or illumination) to be known.

Chapter 6

Estimating Surface Reflectance and Environmental Illumination

Undoing the image formation process and therefore decomposing appearance into its intrinsic properties is a challenging task due to the under-constrained nature of this inverse rendering problem. Nevertheless, as shown in the previous chapters, significant progress has been made on inferring shape, reflectance or illumination from images only.

In this chapter, we present a method that estimates reflectance and illumination information from a single image, where the input image depicts a single-material object of a given class with a specular material and under natural illumination, directly addressing Research Question 4: *Is it possible to decompose a single image into its intrinsic 3D shape, surface reflectance and environmental illumination and if so, what assumptions should be made to make this decomposition feasible?* In contrast to earlier work - in literature and previous chapters - we follow a data-driven, learning-based approach and do not assume one or more components (shape, reflectance or illumination) to be known. To achieve this, we propose a two-step approach, where we first estimate the object's reflectance map, and then further decompose the latter into reflectance and illumination. For the first step[1], we introduce a CNN

[1]The first step of our two-step approach was originally published in the paper: K. Rematas, T. Ritschel, M. Fritz, E. Gavves and T. Tuytelaars, *Deep Reflectance Maps*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.

that directly predicts a reflectance map from the input image itself, as well as an indirect scheme that uses additional supervision, first estimating surface orientation and afterwards inferring the reflectance map using a learning-based sparse data interpolation technique. For the second step, we suggest a CNN architecture to reconstruct both reflectance parameters (*i.e.* Phong parameters) and illumination (*i.e.* high-resolution spherical environment maps) from the reflectance map. We also propose new datasets to train these CNNs.

This chapter is based on the paper:

- S. Georgoulis[2], K. Rematas[2], T. Ritschel, E. Gavves, M. Fritz, L. Van Gool and T. Tuytelaars, *Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning.* Published in IEEE Journal on Pattern Analysis and Machine Intelligence (PAMI) 2017.

6.1 Introduction

A classic computer vision task is the decomposition of an image into its intrinsic properties, *i.e.* its shape, reflectance and illumination. The physics of image formation is based on the complex interplay of these properties; the light (*i.e.* illumination) hits a surface with specific orientation (*i.e.* shape) and material properties (*i.e.* reflectance) and is reflected to the camera. Factoring an image into its intrinsic components, however, is a very difficult and under-constrained task, as the same visual result might be due to many different combinations of intrinsic object properties.

For the estimation of those properties, a common practice in literature is to assume one or more components to be known or simplified and try to estimate the others. On the one hand, traditional approaches to intrinsic images or shape-from-shading try to constrain either reflectance, by assuming Lambertian materials [10, 168, 12], or illumination, by having a controlled lighting environment such as point light sources [67, 166]. On the other hand, recent approaches allow for less constrained reflectance and illumination. Yet, in this case shape is either assumed to be known, given in the form of a scanned 3D model [93], or it is restricted to having trivial geometry (*e.g.* spheres) [51].

We go beyond these simplifying assumptions and estimate reflectance and illumination in a more general setting where the shape of the object is not given but instead it comes from a known class (*e.g.* cars). This is motivated by the observation that as humans we probably exploit a lot of high-level semantic

[2]S. Georgoulis and K. Rematas contributed equally to this work.

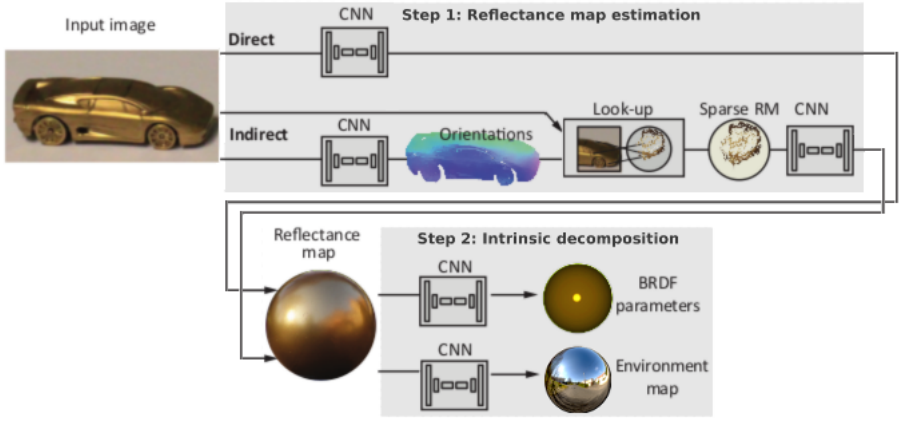


Figure 6.1: Overview of our approach. From the input image, in a first step we estimate a reflectance map either directly from the input image itself or indirectly with additional supervision, and in a second step we decompose the reflectance map into reflectance parameters and an environment map.

cues for similar tasks (*e.g.* car bodies have similar local structures). As such, focusing on objects from a known class allows us to exploit these cues in a learning-based scheme. Furthermore, we observe that there are strong priors about illumination and photo content (*e.g.* the sky is blue and always on top), that are hard to capture in parametric models [126] or carefully designed physics formulas [93]. Instead, going for a data-driven, learning-based approach, allows to naturally embed such priors in the learning process. The latter is essential for dealing with ambiguous cases one encounters in these decomposition problems. We consider this shift from a model-based to a data-driven, learning-based approach as one of the contributions of our work.

To keep the complexity of the problem under control, we propose a two-step approach. First, we estimate a shape-independent representation of the appearance, in the form of a reflectance map [67]. Second, we decompose it into material and illumination. To carry out these tasks we employ CNNs as they have shown unprecedented performance in other de-convolution tasks with similar requirements [38, 37, 168]. Moreover, since reflection has similarities to a convolution of material (*i.e.* reflectance) and illumination [118], it comes natural to use a CNN with de-convolutional layers to invert this process. The input to our method is a single 2D image, depicting a single-material object from a known class, and its segmentation mask. The latter is just used for background removal. The output is the reflectance of the object, expressed as BRDF parameters, and the illumination, expressed as a HDR spherical

environment map. An overview of our two-step pipeline can be seen in Fig. 6.1.

Besides allowing for a better understanding and analysis of 2D imagery, the ability to estimate reflectance maps lends itself to a broad spectrum of applications, including appearance transfer, inpainting and augmented reality, while its further decomposition into reflectance and illumination enables powerful image editing applications, such as material transfer and illumination editing.

As mentioned earlier, we opted for a two-step approach where: (1) we estimate a reflectance map from a 2D image (Fig. 6.1, Step 1), and (2) we decompose it into reflectance and illumination (Fig. 6.1, Step 2). There are several reasons behind this choice: (1) A connected framework trained end-to-end (*i.e.* from images to reflectance and illumination) would require a prohibitively large amount of GPU memory. (2) Even if (1) was possible, there is a lack of large scale databases, especially for reflectance and illumination, needed for training such a connected system. (3) Previous approaches in similar tasks [93] have shown that optimizing in discrete steps (some parameters are kept fixed while estimating the rest) helps in keeping the training process stable.

For the reflectance map estimation, we propose two different approaches: The first approach (Fig. 6.1, Direct) directly estimates a reflectance map from the input image using an end-to-end learning framework based on a CNN with de-convolutions. The second approach (Fig. 6.1, Indirect) leverages additional supervision at training time, to first predict per-pixel surface normals, which are then used to compute sparse reflectance maps from the visible normals of the object. Given the sparse reflectance map, a learning-based sparse data interpolation scheme is introduced to arrive at the final reflectance map.

For the decomposition of the reflectance map into material and illumination, we investigate three different approaches: The first approach independently estimates BRDF parameters and illumination using two different CNN architectures. The second approach jointly estimates both by employing a single CNN that shares the first convolutional layers. Finally, the third approach combines the use of CNNs with classic inverse rendering techniques.

Our key contributions can be summarized as:

- We propose the first deep learning formulation to infer reflectance maps from a 2D image and to further decompose them into material parameters and natural illumination.
- We show new capabilities of CNN architectures, mapping from the image to the directional domain, performing learning-based sparse data interpolation as well as mapping from LDR to HDR data.

- In order to train and evaluate our two-step approach, we provide new datasets that include large scale synthetic data to facilitate the training of deep learning models as well as real data to provide a realistic testing regime.

This chapter is organized as follows: Related work is presented in Sec. 6.2. Next, Sec. 6.3 introduces some basic definitions used throughout the paper. In Sec. 6.4 we present our CNN framework for estimating reflectance maps that we further decompose into reflectance and illumination in Sec. 6.5. Following this, Sec. 6.6 describes the new datasets used for training. Experimental results are reported in Sec. 6.7. Finally, Sec. 6.8 concludes this chapter.

6.2 Related Work

Intrinsics are the individual physical properties that yield a scene's appearance through their interaction [10]. As an example, incoming light reflected on a material's surface in the direction of the observer yields an appearance influenced by the surface's reflectance as well as the scene's illumination. 3D shape is another intrinsic property of the objects in a scene, that also influences appearance through the surface's orientation (normals). Ideally, one can retrieve all these pieces of the appearance jigsaw puzzle separately. In practice, even if one fixes a single component (shape, reflectance, or illumination) by assuming it to be known or by just simplifying it, what one is left with is still a hard decomposition problem for the remaining two components, as we show in the previous chapters. Sometimes one also keeps two of the three intertwined, only retrieving the third as a separate entity.

As making assumptions about one or more of the intrinsics is important to get a handle on the decomposition problem, it is also relevant to better understand their natural statistics. Databases of reflectance [27, 99] or illumination [29, 35] samples have allowed to acquire such statistics, but exploiting them in computation remains challenging. Recent databases focus on images captured in the wild, e.g. annotated for reflectance using crowd-sourcing [12]. We built upon these recent advances and propose a new dataset that captures reflectance maps and normals for the specular case, which are not well represented in prior recordings (*e.g.* the intrinsic image decomposition tasks [102, 168] assume diffuse surfaces).

Next, we describe related work, that we mainly found in the three core research strands listed below. The following discussion gradually homes in on work that gets closer and closer to ours.

Reflectance maps It is not always required to separate reflectance and illumination. *Reflectance maps* [67] - that assign an appearance (i.e. RGB color) to a surface orientation, thereby combining reflectance and illumination - suffice for many important applications. Examples are novel view synthesis (if the 3D shape is available) [123, 121] or material exchanges [73]. Such reflectance maps can be obtained in multiple ways, e.g. using Internet photo collections of diffuse objects to produce a rough 3D shape and then extracting reflectance maps in a second step [56].

In computer graphics, reflectance maps are popular to capture, transfer and manipulate the orientation-dependent appearance of photo-realistic or artistic shading. They are also known as "lit spheres" [137] or "MatCaps" [138]. A special user interface is typically required to map surface orientation to appearance at sparse points in an image, from which orientations are interpolated for in-between pixels to fill the lit sphere (e.g. Rematas *et al.* [123] manually aligned a 3D model with an image to generate reflectance maps). Khan *et al.* [73] made small diffuse objects in a single cluttered image to appear specular or transparent, but they rely on manual interventions and mainly aim for plausible photo-realistic results. Instead, our results do not just look plausible, but stay closer to desired ground-truth even when scene parameters are changed significantly. Surface reflectance and scene illumination are naturally separated in our case.

Factoring Images Classic *intrinsic images* factor an image into reflectance and illumination [10]. Similarly, *shape-from-shading* decomposes into reflectance and shading, eventually leading to an orientation (normal) map or even a full 3D shape.

Recently, factoring images has received renewed interest. Lombardi and Nishino [93] as well as Johnson and Adelson [69] have studied the relation of shape, reflectance and natural illumination. A key idea in these works is, that under natural illumination, appearance and orientation are in a much more specific relation, as used in PS [60], than for a single point light, where many similar appearances for totally different orientations can be present. They present different optimization approaches that allow for estimation of one component if at least one other component is known. In this work, we assume that the object is made of a single material (multi-material objects have to be ruled out), and its object class and segmentation mask are known. The latter is only used to segment the object from the background. We then aim at factoring out reflectance and illumination, in a two-step approach where first we estimate the reflectance map and then we factor the produced reflectance map into material and illumination. As such, our approach solves a more general and less constrained problem compared to approaches such as [93] or [69].

Baron and Malik [9] factor shaded images into shape, reflectance and lighting, but only for scalar reflectance, *i.e.* diffuse albedo, and for limited illumination frequencies. In a very different vein, a recent approach by Richter and Roth [125] first estimates a diffuse reflectance map using approximate normals and then refines the normal map using the reflectance map as a guide. Different from our approach, they assume diffuse surfaces to be approximated using 2nd-order SH and learn to refine the normals from the reflectance map using a regression forest. We compare the reflectance maps produced by our more general approach to reflectance maps using an SH basis, which are limited to diffuse materials only, in our experimental results.

Having estimated the reflectance map from the input image, in the second step of our pipeline we address a problem similar to Lombardi and Nishino [93]: an object with a single, unknown material on the surface (homogeneous surface reflectance) is observed under some unknown natural illumination. Hence, in their case the shape is known (*i.e.* a sphere), and the reflectance and illumination remain to be separately retrieved. Although we address a similar problem as these previous works, our solution is fundamentally different: instead of seeking to invert the physical process under the guidance of manually designed - thus limiting - priors, our work entirely relies on data to learn the backward mapping from a reflectance map to its intrinsics. Our results indicate this inverse mapping can be learned, leading to high-quality, detailed, yet naturalistic environment maps. The underlying network has learned cues such that the fact that windows are bright or that it is the sky that is blue and not so much the object, which can not be modeled by carefully designed physics formulas [93]. Moreover, our approach is the first to perform a slightly altered task, that is much closer to practice, where the image to decompose is captured using a LDR sensor, yet the resulting environment map has HDR as required in re-synthesis tasks. Instead, previous works have typically considered either HDR input [93], which implies the capture of multiple exposures per image making the capturing process rather impractical, or produced only LDR environment maps [126].

Deep learning In recent years *CNNs* have shown strong performance across different domains. In particular, the models for object recognition by Krizhevsky *et al.* [76] and detection by Girshick *et al.* [53] can be seen as a layer-wise encoder of successively improved features. Based on ideas of encoding-decoding strategies similar to auto-encoders, convolutional decoders have been developed [164, 82] to decode condensed representations back to images. This has led to fully convolutional or de-convolutional techniques that have seen wide applicability for tasks where there is a per-pixel prediction target. In [94, 58], this paradigm has been applied to semantic image segmentation, whereas in [34], image synthesis was proposed given object class, view and view transformations as

input and synthesizing segmented new object instances as output. Similarly, Kulkarni *et al.* [77] proposed the *deep convolutional inverse graphics networks* with an encoder-decoder architecture, that given an image can synthesize novel views. In contrast, our approach achieves a new mapping to intrinsic properties - the reflectance map, reflectance and illumination.

Deep lambertian networks [144] apply deep belief networks to the joint estimation of a surface's reflectance, an orientation map and the direction of a single point light source. They rely on Gaussian Restricted Boltzmann Machines to model the prior of the albedo and the surface normals for inference from a single image. In contrast, we address specular materials under general illumination, and further factor into reflectance and illumination.

Another branch of research proposes to use neural networks for depth estimation [38, 86, 89], normal estimation [37, 153], intrinsic image decomposition [102, 168] and lightness [103]. Wang *et al.* [153] show that a careful mixture of deep architectures with hand-engineered models allows for accurate surface normal estimation. Observing that normals, depth and segmentations are related tasks, Eigen *et al.* [37] propose a coarse-to-fine, multi-scale and multi-purpose deep network that optimizes depth, normal estimation and semantic segmentation. Likewise, Li *et al.* [86] apply deep regression using CNNs for depth and normal estimation, whose output is further refined by a conditional random field. Going one step further, Liu *et al.* [89] propose to embed both the unary and the pairwise potentials of a conditional random field in a unified deep network. In contrast to these approaches, our goal is not normals, but rather reflectance and illumination estimation – although our "indirect approach" estimates normals as a by-product.

6.3 Definitions

Before presenting our two-step pipeline we analyze some basic definitions that will be used throughout the chapter in more detail. We begin with the *reflectance map* $L(\omega) \in \mathcal{S}^+ \rightarrow \mathbb{R}^3$ [67], which is a map from orientations ω in the positive half-sphere $\mathcal{S}^+$ to the RGB radiance value L leaving that surface to a distant viewer. It combines the effect of *reflectance* and *illumination.*[3]

There are multiple ways to parametrize orientation ω. Horn and Sjoberg [67] used positional gradients which are suitable for an analytic derivation but less attractive for computation as they are defined on the infinite real line.

[3]For the case of a mirror sphere, as it is here, it captures illumination [29] but is not limited to it. It also does not only capture surface reflectance [27], which would be independent of illumination, but rather joins the two.

We instead parameterize the orientation simply by s, t the normalized surface normal's x and y components. Dropping the z coordinate is equivalent to drawing a sphere under orthographic projection with exactly this reflectance map. Note, that orientations of surfaces in an image only cover the upper half-sphere $\mathcal{S}^+$, so we only need to parametrize a half-sphere, avoiding to deal with spherical functions

To arrive at the notion of reflectance maps, as well as the surface reflectance model that we will use, we recall the definition of the rendering equation [70] (RE) which states that for one wavelength

$$L_{\mathrm{o}}(\mathbf{x}, \omega_{\mathrm{o}}) \quad = \quad L_{\mathrm{e}}(\mathbf{x}, \omega_{\mathrm{o}}) \; + \; \int_{\Omega^+} f_{\mathrm{r}}(\mathbf{x}, \omega_{\mathrm{i}}, \omega_{\mathrm{o}}) L_{\mathrm{i}}(\mathbf{x}, \omega_{\mathrm{i}}) \left\langle \omega_{\mathrm{i}}, n(\mathbf{x}) \right\rangle^{+} \mathrm{d}\omega_{\mathrm{i}}, \quad (6.1)$$

where L_{o} is the outgoing radiance, L_{e} the emitted radiance, L_{i} the incoming radiance, f_{r} the BRDF and $n(\mathbf{x})$ the surface orientation. The radiances are both functions of position $\mathbf{x}$ and direction ω. The reflected part is the integral over the upper hemisphere $\mathcal{S}^+$ of the product of incoming light L_{i}, BRDF f_{r} and the dot product of surface normal $n(\mathbf{x})$ and integration direction ω_{i}. In this work it is assumed that *i)* there is no emission, *ii)* the positions of light entry and exit do not differ (translucent objects are excluded), *iii)* there is only a single material (one surface reflectance model to be considered), *iv)* the object is seen under orthographic projection from an infinitely far-away observer, *v)* that the incoming light comes from a distant scene and as such only depends on direction (environment map illumination), and *vi)* there are no shadows. These simplify the RE to the following function

$$L_{\mathrm{o}}(\omega_{\mathrm{o}}) = \int_{\Omega^+} f_{\mathrm{r}}(\omega_{\mathrm{i}}, \omega_{\mathrm{o}}) L_{\mathrm{i}}(\omega_{\mathrm{i}}) \left\langle \omega_{\mathrm{i}}, \mathbf{n} \right\rangle^{+} \mathrm{d}\omega_{\mathrm{i}}, \quad (6.2)$$

which refers to the reflectance map L_{o} of the illumination L_{i} and the surface reflectance f_{r}. Henceforth, for simplification we refer to the surface reflectance model f_{r} as the *material*. A data-driven BRDF would be an ideal such reflectance model, but here it is simplified to the seven-parameter Phong model [116]

$$f_{\mathrm{r}}(\omega_{\mathrm{i}}, \omega_{\mathrm{o}}) = k_{\mathrm{d}} + k_{\mathrm{s}} \cdot \left\langle r(\omega_{\mathrm{i}}, \mathbf{n}), \omega_{\mathrm{o}} \right\rangle^{k_{\mathrm{g}}}, \quad (6.3)$$

where k_{d} is called the *diffuse color*, k_{s} the *specular color*, k_{g} the *glossiness*, and $r(\cdot, \cdot)$ the mirror reflection of L_{i}.

As both the illumination L_{i} and the reflectance map L_{o} are two-dimensional functions of direction ω, we represent them as images using the described s, t parameterization. Nevertheless, other parameterizations could also be used, such as the Lambert, latitude-longitude or the mirror-ball mappings [30].

6.4 Step 1: From Images to Reflectance Maps

In this section, we present a solution for the first step of our pipeline, which is the estimation of the reflectance map when a single 2D image depicting a single-material object from a known class (*e.g.* cars) and its segmentation mask are given as input.

Motivation We address a challenging inverse rendering problem that is highly under-constrained. Therefore, any solution needs to mediate between evidence from the data and prior expectations. In the general setting of specular materials and unknown natural illuminations, modeling prior expectations over reflectance maps - let alone obtaining a parametric representation - seems problematic. This motivated us to follow a data-driven approach in an end-to-end learning framework, where the dependence of reflectance maps on object appearances is learned from a substantial number of synthesized images for a given object class.

Overview We want to estimate the reflectance map of a single-material object depicted in a single RGB image (see Fig. 6.1, Step 1). This is equivalent to estimating how a sphere [137] with the same material as the object would look like from the same camera position and under the same illumination. We propose two approaches to estimate reflectance maps: a *Direct* (Sec. 6.4.1) and an *Indirect* one (Sec. 6.4.2). Both have a general RGB image as input and a reflectance map as an output. The *Indirect* method also produces a conjoint per-pixel normal map. Both variants are trained from and evaluated on the new SMASHINg dataset introduced in detail in Sec. 6.6.1. For now, we can assume the training data to consist of pairs of 2D RGB images (domain) and reflectance maps (range) with the latter in the parameterization explained in Sec. 6.3.

6.4.1 Direct Approach: An End-to-End Learning-Based Model for Inferring Reflectance Maps

In the *Direct* approach (Fig. 6.1, Step 1, Direct), we learn a mapping between the object's segmented image and its reflectance map, following a convolutional-deconvolutional architecture.

Fig. 6.2 shows the proposed architecture. Starting from a series of convolutional layers, each followed by batch normalization, ReLU and pooling layers, the size of the input feature maps is reduced to 1×1. After continuing with two

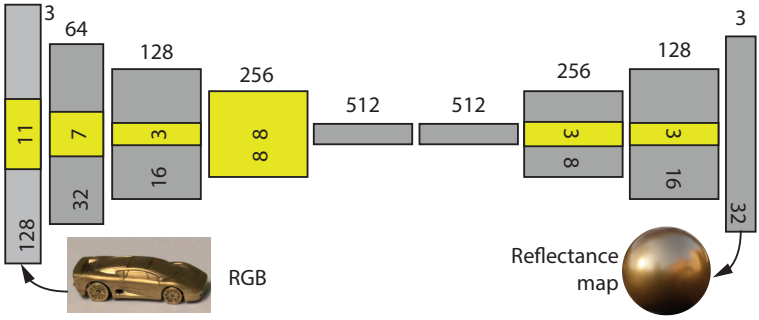


Figure 6.2: Architecture of the *Direct* approach for the reflectance map estimation (see also Fig. 6.1, Step 1, Direct). The bottom numbers represent the spatial resolution and the ones on top the size of the feature channels for the corresponding convolutional layer. Finally, the yellow boxes in the middle indicate the filters' size.

fully-connected layers, the feature maps are up-sampled until the output size is 32×32 pixels. In all convolutional layers a stride of 1 and zero padding are used such that the output has the same size as the input. The final layer uses an Euclidean loss between the RGB values for the predicted and the ground-truth reflectance map.

In short, for the *Direct* approach the network needs to learn how to "encode" the input image to a reflectance map. Note that this is a particularly challenging task, as the model has to learn not only how to map the image pixels to locations on the reflectance map (change from image to directional domain), but also to impute and interpolate appearance for unobserved normals.

6.4.2 Indirect Approach: Estimating Reflectance Maps from Inferred Normals Using Sparse Data Interpolation

As an alternative for the *Direct* approach described above, we also explored an *Indirect* approach, that explicitly incorporates domain knowledge about the RE and the relation between the input image and corresponding reflectance map.

The *Indirect* approach (Fig. 6.1, Step 1, Indirect) proceeds in four steps: *1a)* estimating per-pixel orientation maps from the RGB image, *1b)* up-sampling the orientation map to the full available input image resolution, *1c)* changing from the image domain into the directional domain, producing a sparse reflectance map, and *1d)* predicting a dense reflectance map from the sparse one.

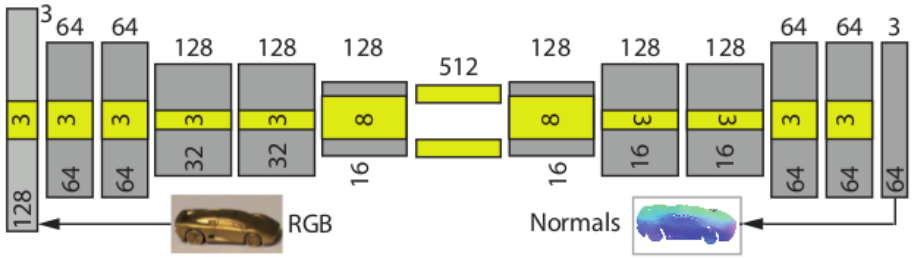


Figure 6.3: Architecture of the normals estimation sub-step of our *Indirect* approach for estimating the reflectance map (see also Fig. 6.1, Step 1, Indirect). The notation is the same as in Fig. 6.2. Note, that the middle elements here correspond to fully convolutional filters.

The steps *1a* and *1d* are modeled by CNN architectures, while the steps *1b* and *1c* are prescribed transformations, related to the parametrization of the reflectance map. Next, we detail each step.

(1a) Normals estimation Our goal is to predict a surface orientation (normal) map from the RGB image. To this end, we use our simplified parameterization of the directional domain to coordinates in a flat 2D image of a lit sphere (see Sec. 6.3). Specifically, we seek to find the s, t parameters according to our reflectance map parameterization.

For this task we train a CNN, whose architecture is shown in Fig. 6.3. Inspired by recent works in normals estimation from a single image [37, 153, 86], we opted for a deeper architecture which has proven more efficient in this task. Specifically, the network is fully convolutional as in [94] and it consists of a series of convolutional layers, each followed by ReLU and pooling layers, that reduce the spatial extent of the feature maps. After the fully convolutional layers, there is a series of de-convolutional layers that up-scale the feature representation to half of the original image's size. Finally, we use two Euclidean losses between the predicted and the L_2 normalized ground-truth normals. The first one takes into account only the s, t coordinates of our simplified parametrization, while the second uses the original x, y, z coordinates of the normals (also explaining why the features channel in the last layer of Fig. 6.3 has a size of 3 instead of 2). We have experimentally found that this estimation helps in improving the quality of predicted normals.

(1b) Normals up-sampling In the above network the orientations are estimated at a decimated resolution of 64×64, so the number of orientation samples is in the order of thousands. The input images however are of resolution

128×128 with ten-thousands of pixels. Note that, a full-resolution orientation map is useful for resolving all appearance details in the orientation domain. Also, the appearance of one orientation in the reflectance map can be related to all high-resolution image pixels with that orientation. As such, intended applications performing shape manipulation in the 2D image (cf. Sec. 6.7.3) will benefit from a refined map. To produce this high-resolution orientation map, we use joint bilateral upsampling [75] as also done in range images [18].

(1c) Change-of-domain Next, we want to reconstruct a sparse reflectance map from the high-resolution orientation map of the previous step and the input image. This is a prescribed mapping transformation: The pairs of appearance L_p and orientation ω_p in every pixel p are unstructured samples of the continuous reflectance map function $L(\omega)$ $(= L_o(\omega_o))$ we seek to recover. Our goal now is to map these samples from the image to the directional domain, constituting the reflectance map. The most straightforward solution is to perform scattered data interpolation

$$L(\omega) = (\sum_{p=1}^{n} w(\langle \omega, \omega_p \rangle))^{-1} \sum_{p=1}^{n} w(\langle \omega, \omega_p \rangle) L_p, \tag{6.4}$$

where $w(x) = \exp(-(\sigma \cos^{-1}(x))^2)$ is an RBF kernel.

In practice however, the orientation estimates are noisy and the requirements of a global reflectance map (directional illumination, orthographic projection, no shadows) are never fully met, asking for a more robust estimate. We found darkening due to shadows to be the largest issue in practice. Therefore, we instead perform a max operation over all samples closer than a threshold $\epsilon = \cos(5°)$,

$$L(\omega) = \max\{w(\langle \omega, \omega_p \rangle) L_p\}, \quad w(x) = \begin{cases} 1 & \text{if } x > \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

If one orientation is observed under different amounts of shadow, only the one that is not in shadow will contribute - which is the intended effect. Still, the map resulting from this step is sparse due to normals that were not observed in the image, as seen in the example of Fig. 6.4 (Sparse RM). This requires imputing and interpolating the sparse data in order to arrive at a dense estimate.

(1d) A learning-based approach for sparse data interpolation The result of the previous step is a sparse reflectance map, that is noisy due to errors from incorrectly estimated normals and has missing information at orientations that were not observed in the image. Note, that the latter is not a limitation of

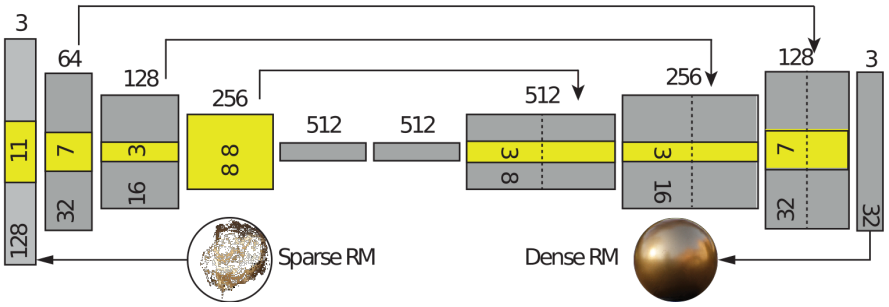


Figure 6.4: Architecture of the last sub-step of our *Indirect* approach for the reflectance map estimation. The notation is the same as in Fig. 6.2.

the normal estimation, but even occurs for ground-truth surface orientations: If an orientation is not present, its appearance remains unknown.

One solution is to directly use Eq. 6.4 to get a dense output. As will be shown in Sec. 6.7.1 though, this leads to poor performance. Instead, we propose a learning-based approach to predict a dense reflectance map from a sparse and noisy one. Accordingly, the network is trained on pairs of sparse and dense reflectance maps. The sparse ones are created using the steps *1a-1c* on synthetic data where the target reflectance map is known by rendering a sphere.

The employed CNN architecture is shown in Fig. 6.4. Note, that it is very similar to the architecture of our *Direct* approach (see Fig. 6.2). Input is the sparse reflectance map and output the dense one. Since both are in the same domain, we use the output of the convolutional layers as additional cues. Specifically, after each de-convolution layer, we concatenate its output with the feature map from the respective convolution layer. This is a common practice in CNN architectures with similar tasks as it helps preserving the local structure of the predicted image. Finally, an L_2 loss between the predicted and the ground-truth dense reflectance map is used.

6.5 Step 2: From Reflectance Maps to Material Parameters and Natural Illumination

In the previous section we presented our approach for estimating the reflectance map of an object from a single image. In what follows, we show how to further decompose the estimated reflectance map into its intrinsics: material (*i.e.* reflectance) and illumination.

Overview The input to our material and illumination decomposition (Fig. 6.1, Step 2) is the LDR reflectance map estimated from the first stage of our pipeline (Fig. 6.1, Step 1). In general, a reflectance map can be obtained in several other ways. For example, when a spherical sample of the desired material is available, it can directly be put under the desired illumination and photographed. In practice, this is usually not the case though - the sample has a different shape. If the shape is known, *i.e.* its normals are known, its reflectance map can be retrieved, at least for all observed surface orientations. In this latter case, only the last step of our indirect reflectance map estimation needs to be applied. If the shape is unknown, several options have been explored to acquire it, including 3D scanning, SfM, depth sensors, CNN-based depth extraction [38, 86, 89] or directly estimating the normals using deep learning [37, 153]. Although in this paper we assume that the reflectance map is given from the first stage of our pipeline (Fig. 6.1, Step 1), for the sake of generality it is useful keeping these other options in mind.

The outputs of Step 2 are: (1) the Phong reflectance parameters (see Eq. 6.3) and (2) an HDR environment map in the parameterization of Sec. 6.3. The environment map is an HDR spherical image, expressing illumination's directional dependency. Remember that HDR is a critical property to have for illumination [30, 35], as without it re-illumination is likely to fail in many real-world cases. Note that our estimated illumination is still HDR even when the input is only LDR, which is a generalization over previous approaches that require HDR inputs [93, 92].

We enable this mapping by proposing *DeLight-Net*, a framework of CNNs, trained on synthetic data. All CNNs take as input a dense reflectance map. *Material CNN* outputs a parameter vector, which is 7-dimensional in the case of the Phong reflectance model: one color for the diffuse, one for the specular component, and a glossiness value, defining how shiny the material is (see Eq. 6.3). *Illumination CNN* outputs the HDR environment map. These two independent CNNs comprise our INDEP. approach. We also propose two variants: JOINT shares intermediate representations to perform the estimation jointly, and SEQUEN. combines the *Illumination CNN* with classic inverse rendering techniques to estimate the Phong parameters.

6.5.1 Independent Material and Illumination Estimation

Our INDEP. approach builds on *Material CNN* and *Illumination CNN* to independently estimate material parameters and natural illumination from a dense reflectance map. For both networks we used Huber loss for regression.

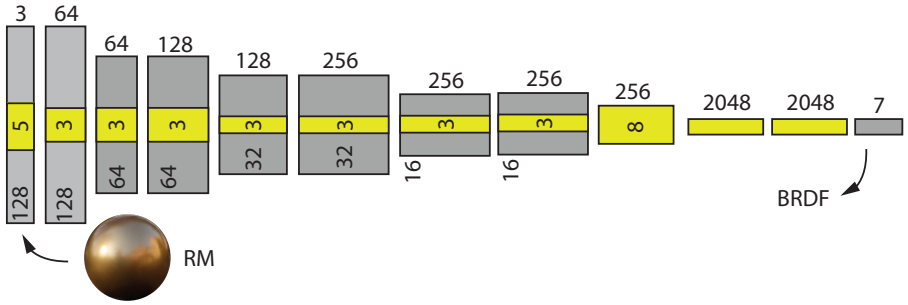


Figure 6.5: The *Material CNN* for estimating Phong reflectance parameters. The notation is the same as in Fig. 6.2.

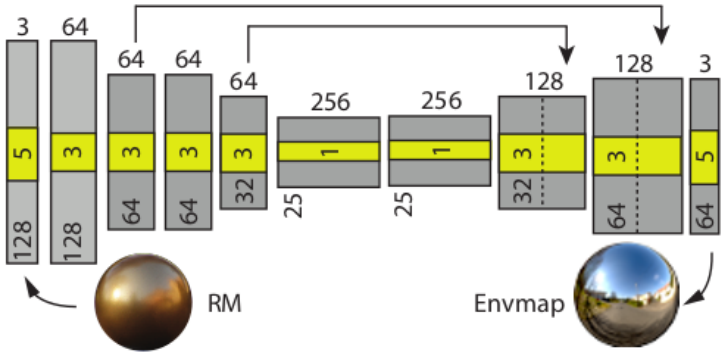


Figure 6.6: The *Illumination CNN* for estimating natural illumination. The notation is the same as in Fig. 6.2.

We have experimentally found that this choice nicely balances between learning the dynamic range and the color distribution of the environment map.

Material CNN As already mentioned, the input to this network is a 2D image of the dense reflectance map, while the output is a 7-parameter Phong vector. The design of the network is shown in Fig. 6.5. Overall, the network consists of multiple convolutional layers reducing the resolution, followed by several fully-connected layers. Note that each convolutional unit is always followed by batch normalization and ReLU.

Illumination CNN As already mentioned, the input to this network is the same dense reflectance map as in *Material CNN*, while its output is an HDR

environment map of half the spatial resolution (Fig. 6.6). The feature spatial resolution is gradually reduced by about one order of magnitude, from 128 to 25, with the middle layers applied in a fully convolutional fashion. Also, two layers of de-convolution are added, that take intermediate results from the previous same-resolution convolutional layers into account (similar design as in Fig. 6.4). We remind that by doing so, fine spatial details can be preserved. As before, each convolutional unit is always followed by batch normalization and ReLU.

6.5.2 Joint Material and Illumination Estimation

Besides the independent estimation of material and illumination, as discussed in Sec. 6.5.1, we also experimented with a network estimating both somewhat more jointly. In this JOINT approach, the network shares the first two layers of *Material CNN* and *Illumination CNN*, as seen in Fig. 6.5 and Fig. 6.6 respectively, and is consequently split, preserving the individual architectures that result in two outputs with their independent losses.

6.5.3 Sequential Material and Illumination Estimation

While the two approaches explained above can estimate material parameters and natural illumination separately or jointly, we also investigate an alternative that combines CNNs with classic inverse rendering. For this SEQUEN. approach, we use the output of the *Illumination CNN* as an input to a classic inverse rendering solution for material estimation. To this end, we show how Phong reflectance parameters can be estimated from a reflectance map and known illumination in a closed form solution. Going back to our simplified reflectance map from Eq. 6.2, when BRDF f_r is Phong,

$$L_\mathrm{o}(\omega_\mathrm{o}) = \underbrace{k_\mathrm{d} \int L_\mathrm{i}(\omega_\mathrm{i}) \left\langle \omega_\mathrm{i}, \mathbf{n} \right\rangle^+ \mathrm{d}\omega_\mathrm{i}}_{\text{Diffuse}} + \underbrace{k_\mathrm{s} \int L_\mathrm{i}(\omega_\mathrm{i}) \left\langle (r(\omega_\mathrm{i}, \mathbf{n}), \omega_\mathrm{o}) \right\rangle^{k_\mathrm{g}} \left\langle \omega_\mathrm{i}, \mathbf{n} \right\rangle^+ \mathrm{d}\omega_\mathrm{i}}_{\text{Specular}}, \tag{6.6}$$

it can be written as a linear combination of a diffuse reflectance map L_d and a gloss-dependent specular reflectance map L_s:

$$L_\mathrm{o}(\omega_\mathrm{o}) = k_\mathrm{d} \underbrace{L_\mathrm{d}(\omega_\mathrm{o})}_{\text{Diffuse RM}} + k_\mathrm{s} \underbrace{L_{\mathrm{s},k_\mathrm{g}}(\omega_\mathrm{o})}_{\text{Specular RM}} .$$

Having observed many pixel samples of L_o, and having estimated L_i using *Illumination CNN*, L_d and $L_{\mathrm{s},k_\mathrm{g}}$ can be computed for all values of k_g.

Figure 6.7: Our dataset for the reflectance map estimation consists of synthetic images with random view, 3D shape, material, illumination and exposure.

Furthermore, if we hold k_g fixed, estimating k_d and k_s is a linear least-squares problem: Let $\mathbf{l}_\mathrm{o}, \mathbf{l}_\mathrm{d}, \mathbf{l}_{\mathrm{s},k_\mathrm{g}}$ be vectors of those pixels for a gloss level k_g. So $\mathbf{l}_\mathrm{o} = k_\mathrm{d}\mathbf{l}_\mathrm{d} + k_\mathrm{s}\mathbf{l}_{\mathrm{s},k_\mathrm{g}}$ or $\mathsf{A}\mathbf{x} = b$, where $\mathsf{A} = (\mathbf{l}_\mathrm{d}|\mathbf{l}_{\mathrm{s},k_\mathrm{g}})$, $\mathbf{x} = (k_\mathrm{s}, k_\mathrm{d})$, and $\mathbf{b} = \mathbf{l}_\mathrm{o}$. This can efficiently be solved for $\mathbf{x}$ for every gloss level k_g by inverting a 2×2 matrix. In order to find the optimal gloss level k_g, a line search for discrete gloss levels is performed, in our case on 100 levels, logarithmically spaced.

This procedure is only applicable because the number of non-linear parameters is low in the Phong model and would not scale to more complex material models. Still, as we show later, estimating Phong parameters analytically and illumination using *Illumination CNN* is outperforming more complex material models.

6.6 Datasets

To train the two steps of our pipeline a large number of images is required. Since it is very difficult to acquire many real images - at least in the order of ten-thousands - together with their ground-truth 3D shape, material (*i.e.* reflectance) and HDR illumination, we opted for synthetically rendered images for the training process. Unfortunately, there is also a lack of large scale databases of scanned material samples and HDR environment maps. As such, we generated two datasets for training each step of our pipeline with emphasis given on different aspects every time.

6.6.1 The SMASHINg challenge dataset

For the reflectance map estimation (Fig. 6.1, Step 1), we propose the Specular MAterials on SHapes with complex IllumiNation (SMASHINg) challenge. It includes a dataset of real as well as synthetic images, ground-truth reflectance maps and normals (where available), results from different methods for baseline comparisons and a set of metrics that we propose to evaluate and compare performance. The data, baselines, methods as well as performance metrics are publicly available[4].

Our dataset combines synthetic images, photographs and images from the web, all depicting cars. We have manually segmented foreground and background for every image.

Synthetic images Synthetic images are produced with random *i)* views, *ii)* 3D shapes, *iii)* materials, *iv)* illumination and *v)* exposure. A preview can be seen in Fig. 6.7. The view is sampled from a random position around the object, looking at the center of the object with a FOV of 40°. The 140 3D shapes come from the free 3D Warehouse repository, indexed by Shapenet [22]. For each sample the object orientation around the y axis is randomized. Illumination is provided by 40 free HDR environment maps collected from the Internet (for more details visit the project's webpage). The exposure is sampled over the "key" parameter of Reinhard et al.'s photographic tone mapper [120], between 0.4 and 0.6. For materials, the MERL BRDF database [99] containing 100 materials is used. Overall 60 k sample images from that space are generated. We define a training-test split so that no shape, material or illumination is shared between the training and test set.

Photographs As real test images, we have recorded photos of six toy cars that were completely painted with a single car lacquer, placed in four different lighting conditions and photographed from five different views, resulting in a total of 120 images. For the corresponding ground-truth reflectance maps, we placed in the same locations spheres painted with the same material. Again, those real images were manually segmented from the background.

Internet images In order to provide an even more challenging test set, we collect an additional 32 car images from the Internet. Here we do not have access to ground-truth normals or reflectance maps, but this setting provides a realistic test case for imaged-based editing methods. Again, we have manually

[4]Project webpage: `https://homes.cs.washington.edu/~krematas/DRM/`

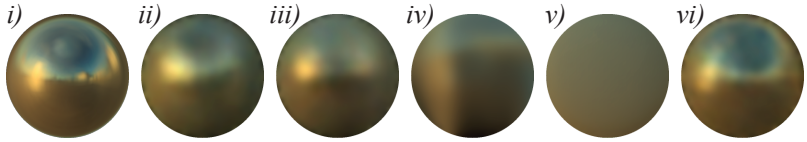


Figure 6.8: Different methods to reconstruct reflectance maps.

segmented out the car body. This allows the qualitative evaluation of single-material normal and reflectance map prediction. Note that, for the Internet images we used networks that were trained on synthetic data from segmented meshes to contain only the car body.

Methods and metrics We include six different methods to reconstruct reflectance maps, visualized in Fig. 6.8: *i)* ground-truth, *ii)* our *Direct* approach, *iii)* our *Indirect* approach, *iv)* an approach that follows our *Indirect* one, but instead of using a CNN for sparse interpolation, it relies on an RBF reconstruction as described in Eq. 6.4, *v)* SH where we project the ground-truth reflectance map to the SH domain, and *vi)* an *Indirect* approach where the estimated normals are replaced by ground-truth normals.

To assess the quality of the reflectance map estimation step we employ two different metrics. The first is the traditional L_2 error between all defined pixels of the reflectance map in RGB and the second is the DSSIM structural difference [155] that excels in measuring the similarity between two images.

6.6.2 DeLight-Net Dataset

For the reflectance map decomposition (Fig. 6.1, Step 2) our training data consist of a set of synthetically rendered images of reflectance maps with random materials under random HDR illuminations from random views. Fig. 6.9 shows such examples of training data.

Training materials were again taken from the MERL BRDF database [99], in particular, the Phong fit made therein. There are 100 materials overall - in our case 67 were used for training and 33 for testing - including diffuse, glossy, and mirror-like appearances.

For illumination we used 70 free HDR environment maps in total - 60 for training and 10 for testing - from the commercial content supplier HDR Maps (`https://www.hdrmaps.com/`). These images are radiometrically calibrated, *i.e.* they differ from the true physical RGB radiance units by only a factor.

We found this not to be the case for other HDR environment maps found on the Internet, which is crucial for re-lighting. All environment maps were also rendered as mirror spheres and consequently re-sampled to 128×128 pixels, which is the resolution of the maps we will later infer.

View positions are sampled from a random direction in the xz-plane with a random declination of $\pm 10°$; an orthographic projection is used. The shape is always a highly-tessellated sphere with analytic normals. For rendering we use the full convolution of the environment map with the Phong parametric model (see Eq. 6.2). This convolution is computationally demanding and to keep it tractable when producing massive training data, it was implemented using GPUs. The rendering result is a 128×128 image. Overall, we produced approximately 50 k sample images of synthetically rendered reflectance maps. Note that for the testing set both the material as well as the environment map are never seen before. The training and benchmark data as well as the CNN architectures used are made publicly available[5].

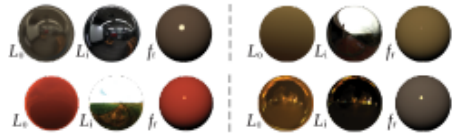


Figure 6.9: Four examples of training data for decomposing reflectance maps into material parameters and natural illumination: Each triplet is a sample of the training set.From left to right: the input reflectance map L_o, the output environment map L_i and material f_r.

Two variants of the resulting images are kept, with slightly different purposes: an HDR and an LDR variant. For the HDR variant we apply the natural logarithm to the RGB data, stored as a 32-bit float image file, as also done in [99] to avoid bias towards differences in the higher intensity ranges during training. For the LDR variant we simulate the exposure process, as follows: First we automatically choose an exposure level using the (5,95)-percentiles. Second, linear radiance values are mapped into the (0,1)-range and quantized uniformly into 256 values (8-bit). Finally, the values are mapped back to absolute radiance and stored in a 32-bit float format. This procedure simulates the information available to a contemporary capturing device with EXIF information (aperture, exposure time, ISO): radiance quantized to 8-bit in an appropriately chosen exposure, allowing to re-scale it to absolute radiance, but with quantization and clipping. In both variants we convert from RGB to CIE LAB color space.

[5]Project webpage: `http://homes.esat.kuleuven.be/~sgeorgou/DeLight/`

6.7 Experiments

In this section, we perform the experimental analysis of the proposed pipeline. Since we rely on a two-step approach, we find it fit to first evaluate each step individually and compare with their corresponding baselines before assessing the system as a whole. As such, in Sec. 6.7.1 we first evaluate the proposed end-to-end *Direct* approach for the reflectance map estimation (Fig. 6.1, Step 1) on the new SMASHINg challenge and compare it to the *Indirect* approach in its different variants (see Sec. 6.6.1). Second, in Sec. 6.7.2 we perform extensive evaluations for the results produced by the reflectance map decomposition framework (Fig. 6.1, Step 2) and compare it with s-o-t-a approaches [90, 93]. Finally, in Sec. 6.7.3 we analyze the qualitative performance of our combined pipeline through various applications including a wide range of image-based editing tasks.

6.7.1 Evaluation of Reflectance Map Estimation

Setup Here, we provide results for our *Direct* method that learns to predict reflectance maps directly from the input image in an end-to-end scheme, as well as several variants of our *Indirect* approach that utilizes intermediate results facilitated by additional supervision through normals at training time (cf. Fig. 6.1, Step 1). The variants of the *Indirect* scheme are based on our estimated normals, but differ in their second stage that has to perform a type of data interpolation to arrive at a dense reflectance map, given the intermediate sparse estimate. For the interpolation, we investigate the proposed learning-based approach, *Indirect (CNN)*, as well as using RBF interpolation, *Indirect (RBF)*. Furthermore, we provide best case analysis by using ground-truth normals in the *Indirect* approach, *Indirect (GT Normals)* (only possible for synthetic data), and computing a diffuse version of the ground-truth by means of SH, *SH (GT Normals)*. The latter gives an upper bound on the result that could be achieved by methods relying on a diffuse material

Table 6.1: Quantitative results for the reflectance map estimation (cf. Fig. 6.1, Step 1) using the different methods defined in Sec. 6.6.1.

	Synthetic		Real	
Method	MSE	DSSIM	MSE	DSSIM
Direct	.0019	.0209	.0120	.0976
Indirect (RBF)	.0038	.0250	.0116	.0814
Indirect (CNN)	.0018	.0180	.0143	.0991
SH (GT Normals)	.0044	.0301	.0114	.0914
Indirect (GT Normals)	.0008	.0111	—	—

assumption. Quantitative results for the different approaches are summarized in Tbl. 6.1.

Reflectance map analysis Overall, we observe consistency among the two investigated metrics, MSE and DSSIM (as defined in Sec. 6.6.1), in how they rank approaches. We obtain accurate estimations for the synthetic set of the SMASHINg challenge dataset for our *Direct* as well as the best *Indirect* methods. The quantitative findings are underpinned by the visual results, e.g. showing the predicted reflectance maps in Fig. 6.10. The performance on the real images is generally lower with the error roughly increasing by one order of magnitude. Yet, the reconstruction still preserves rich specular structures and gives a truthful reconstruction of the represented material.

In more detail, we observe that the best *Direct* and *Indirect* approach perform similar on the synthetic data, although *Direct* did not use the normal information during training. For the real examples, this form of additional supervision seems to pay off more and the RBF interpolation scheme achieves best results in the considered metrics. A closer inspection to the results though, clearly shows the limitations of the image-based metrics. While the RBF-based technique yields a low error, it frequently fails to generate well-localized highlight features on the reflectance map (see also an indicative illustration in Fig. 6.8). We refer the reader to the project's webpage for a detailed visual comparison of all methods.

The ground-truth baselines give further insights into improvements over prior diffuse material assumptions and the future potential of the method. The *SH (GT Normals)* baseline shows that our best methods improve over a best case diffuse estimate with a large margin for the DSSIM metric - highlighting the importance of considering more general reflectance maps. The *Indirect (GT Normals)* illustrates a best case analysis of the *Indirect* approach where we provide ground-truth normals. The results show a potential performance leap by having better estimated normals.

Normals analysis Tbl. 6.2 quantifies the error in the normals estimation by the first stage of our *Indirect* approach. This experiment is facilitated by the synthetic data where normals are available. L_2 corresponds to a network using the Euclidean loss on the x, y, z components of the normals, while *Dual* uses the two losses described in Sec. 6.4.2. *Up*

Table 6.2: Normals estimation of *Indirect* approach on synthetic data.

	Mean	Median	RMSE
L_2	14.3	9.1	20.6
Dual	13.4	8.2	19.8
Dual & Up	13.3	8.2	19.9

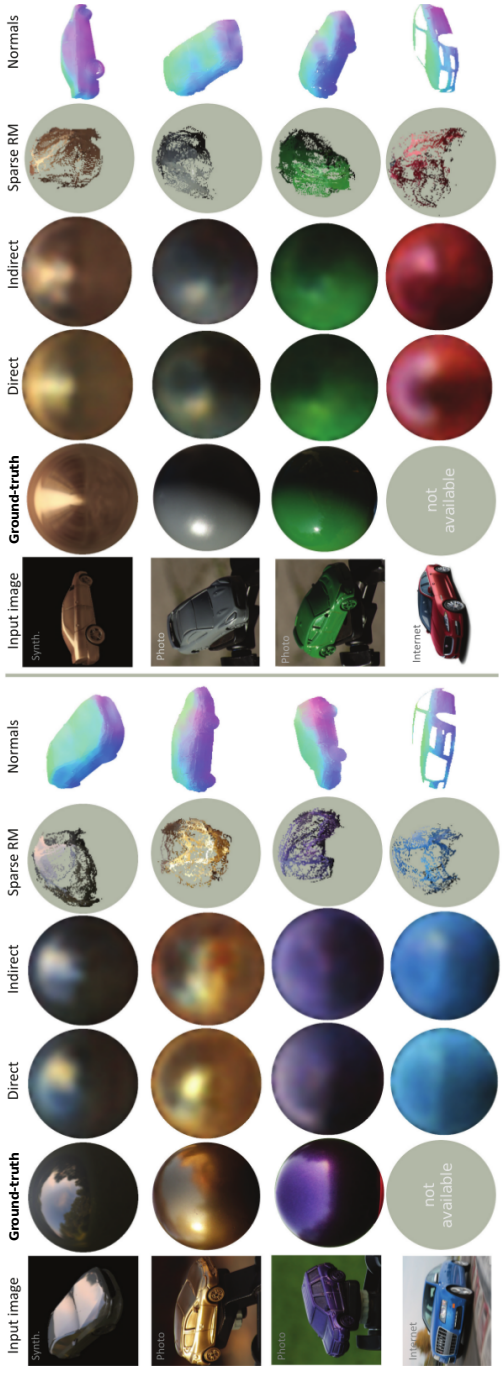


Figure 6.10: Results of different variants and steps for our reflectance map estimation (Fig. 6.1, Step 1).From left to right: input image, ground-truth reflectance map (RM), RM result of the *Direct* approach, RM result of the *Indirect* approach, the intermediate sparse RM produced in the *Indirect* variant, normals produced by the *Indirect* variant as well. Each result is annotated to come from the synthetic, photographed or Internet part of our database. For the Internet-based part, no ground-truth RM is available. We refer the reader to the project's webpage for exhaustive results in this format.

refers to a network trained on up-sampled normals. Both dual loss and joint up-sampling improve the estimated normals. Despite the fact that this analysis is conducted on synthetic data, we observe that our models predict very convincing normals even in the most challenging scenario that we consider (see Fig. 6.10 and Fig. 6.12).

6.7.2 Evaluation of Material and Illumination Estimation

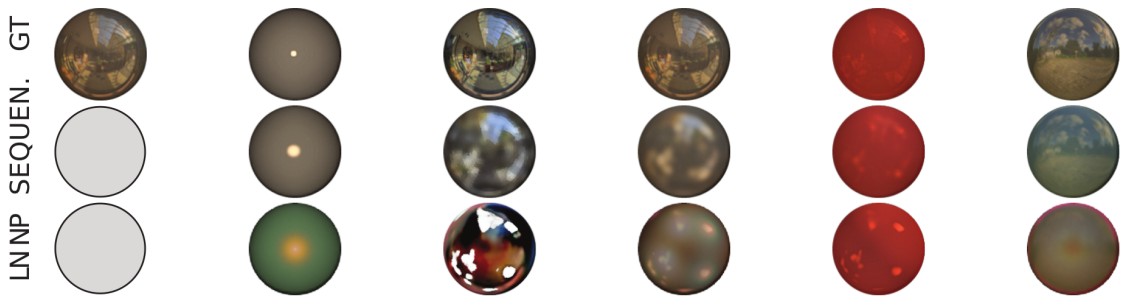
Here, we evaluate our approach for decomposing the reflectance map (cf. Fig. 6.1, Step 2), by first using it in a synthetic re-synthesis benchmark, where images are re-synthesized for original or novel illuminations and materials starting from the estimated components, and second, on real photographs of reflectance maps.

Synthetic re-synthesis benchmark Evaluating a successful decomposition is not trivial due to the complex interaction of material, illumination, shape and viewpoint. The established evaluation protocol [99, 91, 93] is to measure the L_2 error between renderings using the estimated and ground-truth components respectively. Indeed, the reflectance map decomposition into material parameters and natural illumination allows direct evaluation of a) the estimated material parameters by rendering them under a point light source (*Point light*), b) the estimated natural illumination by rendering on a mirror sphere (*Mirror Mat.*), c) both estimated material parameters and natural illumination by re-rendering them together as a reflectance map (*Re-synthesis*).

We enhance this protocol, by also including two extensions inspired by real-world applications: d) we evaluate how well the estimated material parameters perform under different illumination (*Nat. Illum.*). To do so, we compute the reflectance map of the estimated material illuminated by a new environment map, not included in the training set. And finally e) we measure how well the estimated natural illumination generalizes to new materials (*MERL Mat.*). This is performed by selecting a random MERL material, not contained in our training data, and rendering it under the estimated illumination.

The different approaches, represented as rows in Tbl. 6.3, are: "Indep." is our approach with independent CNNs for estimating the material parameters and natural illumination (see Sec. 6.5.1). "Joint" is our joint material and illumination estimation (see Sec. 6.5.2). "Sequen." refers to sequentially estimating natural illumination and material parameters (see Sec. 6.5.3). "LN" refers to the method of Lombardi and Nishino [90]. A comparison with their work is made, both when using the default values for their priors ("LN DP") and when using no priors ("LN NP"), which might depend on the types of materials and illuminations used [90].

Table 6.3: Synthetic evaluation for our material and illumination estimation (Fig. 6.1, Step 2). Rows represent the different approaches and columns the different tasks. Results are reported for two error metrics, LRMSE and DSSIM (lower is better). The images on top are samples from selected rows and columns. The best method for a task is shown in bold.



	Point light		Mirror Mat.		Re-synthesis		MERL Mat.		Nat. Illum.	
	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM
HDR input										
INDEP.	**.0055**	.0677	.0603	.1821	.0118	.0685	.0232	.0341	.0006	.0466
JOINT	.0082	.0753	**.0590**	**.1782**	.0117	.0770	**.0200**	**.0339**	**.0006**	.0529
SEQUEN.	.0062	**.0326**	.0603	.1821	.0016	**.0175**	.0232	.0341	.0008	**.0209**
LN DP [90]	.0245	.1450	.2537	.3299	.0002	.1485	.0288	.0854	.0019	.1423
LN NP [90]	.0263	.1664	.2862	.3124	**.0001**	.0243	.0292	.0433	.0018	.0605
LDR input										
INDEP.	.0082	.0691	.0626	.1901	.0011	.0624	.0270	.0354	.0006	.0472

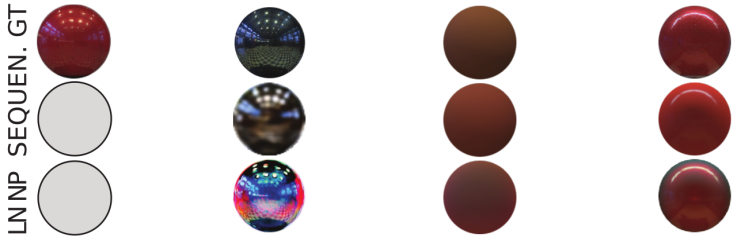
All the above are evaluated on the DeLight-Net dataset (cf. Sec. 6.6.2). Since the method of Lombardi and Nishino [90] require HDR inputs, we perform the comparison on the HDR variant of the dataset (upper part of Tbl. 6.3). We also compare to our INDEP. approach trained on the LDR variant (last row of Tbl. 6.3), which better relates to the LDR reflectance map estimated in the first step of our pipeline. To the best of our knowledge, our method is the first to learn this LDR to HDR mapping.

The final quantitative measure is the difference between the re-synthesized image produced using the estimated decomposition and the re-synthesized image produced using the ground-truth decomposition. The root mean square-error of the logarithm of HDR radiance (LRMSE) and a colored, multi-resolution structured similarity index [155] ran on the tone-mapped LDR result (DSSIM) are used to compare the re-synthesized image to the ground-truth.

Overall, we find that our methods outperform the method of Lombardi and Nishino [90], according to all metrics, with one exception, which is discussed below. When using our estimated material parameters and re-rendering with a point light (*Point light*), our CNNs outperform competitors by a large margin in LRMSE (three-fold improvement) and our SEQUEN. approach by a similar factor according to DSSIM. Using the estimated natural illumination and re-rendering on a mirror sphere (*Mirror Mat.*), is best done using our JOINT approach, again outperforming competitors by a substantial factor according to both metrics. According to LRMSE, for the task of re-synthesizing the input image with both the estimated material parameters and natural illumination (*Re-synthesis*), the approach of Lombardi and Nishino [90] comes out best. This is to be expected, as their approach specifically seeks to minimize in LRMSE the pair of material and illumination that if re-synthesized give the original input. According to DSSIM however, which likely is a better measure, our SEQUEN. approach works best also for this case. When using the estimated components and re-rendering with a new material (*MERL Mat.*) or illumination (*Nat. Illum.*) from the corpus our JOINT approach performs best for both metrics with one exception. According to DSSIM, for the *Nat. Illum.* task our SEQUEN. approach comes out first. Again, the difference to competitors is the strongest in terms of the *Nat. Illum.* task, where a three-fold improvement is achieved, while for the *MERL Mat.* task the difference is almost twice as good. Remarkably, the decomposition performance from LDR inputs is on par with the HDR case, although the problem is more difficult.

In general, our SEQUEN. approach excels in estimating the material parameters whereas our JOINT approach comes marginally first when estimating the natural illunation. We found the latter marginal improvement to be less important in practice, so our SEQUEN. approach is generally the preferred choice.

Table 6.4: Evaluation on real reflectance maps for our material and illumination estimation (Fig. 6.1, Step 2). The notation is the same as in Tbl. 6.3



	Mirror Mat.		MERL Mat.		Nat. Illum.	
	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM
HDR input						
INDEP.	**0.929**	0.376	0.099	0.062	1.111	0.183
JOINT	0.933	**0.365**	**0.052**	**0.043**	**1.110**	0.186
SEQUEN.	0.929	0.376	0.099	0.062	1.223	**0.106**
LN NP [90]	5.402	0.662	1.722	0.071	3.938	0.187
LDR input						
INDEP.	0.950	0.376	0.092	0.059	1.155	0.214

Real reflectance maps The synthetic re-synthesis benchmark has been evaluated on the basis of a large choice of variations on a large number of reflectance maps, illuminations and materials. Capturing a similar amount of reflectance maps ourselves is in practice not possible, so we opted for a smaller set of pairs of materials and environment maps where the ground-truth illumination was also acquired. In particular we use a set of 25 materials under 4 different natural illuminations that we have acquired specifically for this task (see also the project's webpage).

The results are summarized in Tbl. 6.4. The tasks are similar to the ones in our synthetic re-synthesis benchmark, but in a more restricted way, as we do not have the ground-truth material available; such a task would require a gonioreflectometer. As the ground-truth HDR illumination is available however (*i.e.* we scanned it using a chrome sphere), we can compute the difference between the ground-truth illumination and the estimated illumination rendered in a mirror (*Mirror Mat.*). Furthermore, we can re-synthesize, using not just a mirror, but instead a new material from a database, here again MERL (*MERL Mat.*). Finally, we can predict how the estimated material would look under a different illumination, as the same reflectance maps were captured under this different illumination too (*Nat. Illum.*). Note that without ground-truth for the material, re-rendering under point light illumination (*Point light*) and

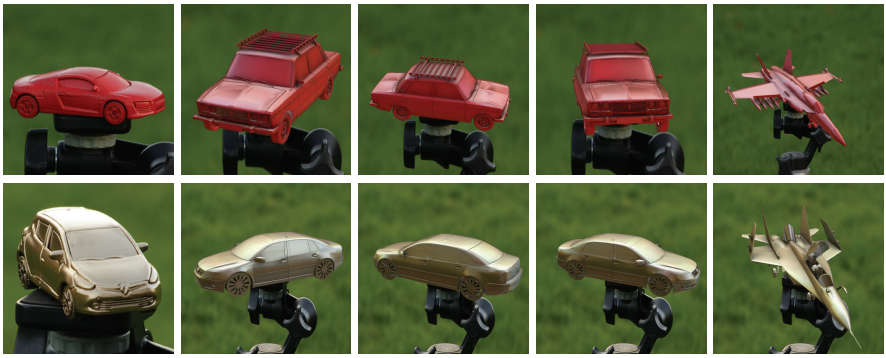


Figure 6.11: Transfer of reflectance maps estimated from real photographs *(1st column)* to virtual objects *(other columns)* of different shape. The project's webpage shows animations of those figures.

re-synthesis using the estimated components (*Re-synthesis*) are not possible. For brevity, the LN NP [90] is compared to our approaches: INDEP., JOINT. and SEQUEN..

We find that the results are consistent with the synthetic evaluation, and our SEQUEN. approach outperforms LN NP [90] for both LRMSE and DSSIM metrics.

6.7.3 Qualitative Results and Applications

Automatically extracting reflectance maps from images - together with the normal information we get as a by-product - and decomposing them into material and illumination facilitates a range of image-based editing applications, such as material acquisition and transfer and shape manipulation. In what follows, we evaluate the performance of our two-step pipeline through numerous example applications. The project's webpages contain all the images and videos that complement our following presentation.

Estimating reflectance maps and normals from images Results of estimated reflectance maps are presented in Fig. 6.10, also showing the quality of the predicted normals. The first row shows two examples on synthetic images, the second and third row on real images and the last row on web images (no reference reflectance map is available here). Notice how the overall appearance, reflecting the interplay between material and the complex illumination, is captured by our estimates. In most examples, highlights are reproduced and

Figure 6.12: Appearance transfer application. Images on the diagonal are the original input. Off-diagonal images have the appearance of the input in their column transferred to the input shape of their row.

even a schematic structure of the environment can be seen in the case of very specular materials.

Inserting virtual objects in a scene Fig. 6.11 shows synthesized images (column 2-5) that we have rendered from 3D models using the reflectance map automatically acquired from the images in column 1. Here, we use ambient occlusion [170] to produce virtual shadows. This application shows how material representations can be acquired from real objects and transferred to a virtual object. Notice, how the virtual objects match in material, specularity and illumination to the source image on the left.

Transferring appearance between images A useful application of our approach is the appearance transfer between different objects in different scenes. To do so, we first estimate reflectance maps for each object independently, swap the estimated reflectance maps, and then use the estimated normals to re-render the objects using a normal look-up table from the new reflectance map. To preserve details, such as shadows and textures, we first re-synthesize each object with its original reflectance map, save the per-pixel difference in LAB color space, then re-synthesize with the swapped reflectance map and add the saved difference in LAB color space. An example is shown in Fig. 6.12. Despite the

Figure 6.13: Shape manipulation application.A user has drawn to manipulate the normal map extracted from our *Indirect* approach. The reflectance map and the new normal map can be used to simulate the new shape's appearance. For a live demo visit the project's webpage.

uncontrolled illumination conditions, we achieve photo-realistic transfer of the appearance, making it hard to distinguish source from target.

Manipulating shape Since we estimate reflectance maps and surface normals, this enables various manipulation applications that work in the directional or normal domain. Fig. 6.13 shows such an application, where the surface orientation is changed, *e.g.* using a special painting interface, and new appearance for the new orientations is sampled from the reflectance map. As before, we save and restore the delta between the original and re-synthesized reflectance map to keep details and shadows. The final result gives a strong sense of 3D structure while maintaining an overall consistent appearance w.r.t. material and scene illumination.

Estimating material and illumination from reflectance maps Starting from a reflectance map we can decompose it into its intrinsic material and illumination. The estimated material parameters and natural illumination can then be used to re-render the object of interest in different scenes, change its material or even replace the object itself with another. Some of the many editing possibilities that our method enables are summarized in Fig. 6.14. Our pipeline allows re-rendering objects with different materials *(horizontal*

variation, Fig. 6.14), under different illuminations *(intra-block vertical variation, Fig. 6.14)*, or for different shapes *(inter-block vertical variation, Fig. 6.14)*. Results are visualized in pairs, where the left half shows re-synthesis using our estimated decomposition and the right half the same re-synthesis using reference material and illumination. The input reflectance maps are marked with a dotted circle. We clearly see that our approach can reconstruct plausible materials and environment maps with fine details.

Manipulating material and illumination from real photos Perhaps the most interesting and practical application is interactive material and illumination manipulation from real photos. We begin from a segmented image of the object of interest, which is the car's body in our case. Using the CNNs of Sec. 6.4 we first estimate the normal orientations and consequently the reflectance map (*Indirect* approach). From the estimated reflectance map we then decompose into material and illumination using the CNNs of Sec. 6.5 (SEQUEN. approach). Finally, we re-render (Fig. 6.15) the imaged object (1*st* column) under different illumination (1*st* row) and different material (2*nd* row). The results for two car models are shown in Fig. 6.15. For more car models you can visit the project's webpage. Note that we have explicitly modeled only the car's body and not the lights, mirrors, windows, etc (same as in Fig. 6.12). The recovered results look nevertheless realistic and convincing.

6.8 Conclusion

We presented a deep learning approach to estimate natural illumination information and surface reflectance characteristics from a single 2D image that facilitates new image-based rendering applications. We show that our technique works with complex 3D shapes, specular materials and under complex natural illumination. In order to achieve our goal, we have developed new deep learning architectures that for the first time achieve sparse data interpolation, mapping from the image to the directional domain, and inferring HDR data from LDR input. The application of deep learning techniques to this domain is facilitated by our novel large scale synthetically rendered dataset that is accompanied by real-world testing data in order to evaluate our approach. Our proposed methods outperform prior work in this area, which highlights the potential of deep learning approaches in inverse rendering tasks and computer graphics in general.

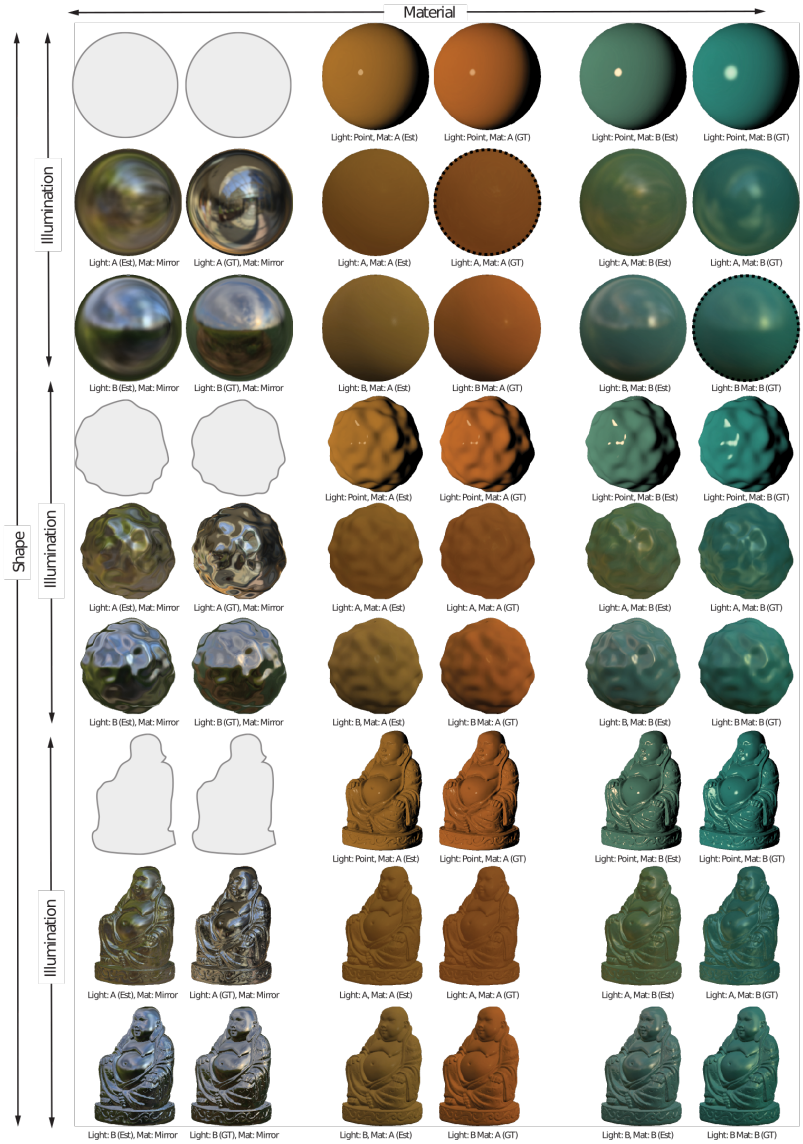


Figure 6.14: Some of the many manipulations enabled by our approach. Please see *Estimating material and illumination from reflectance maps* in Sec. 6.7.3 for a more thorough explanation.

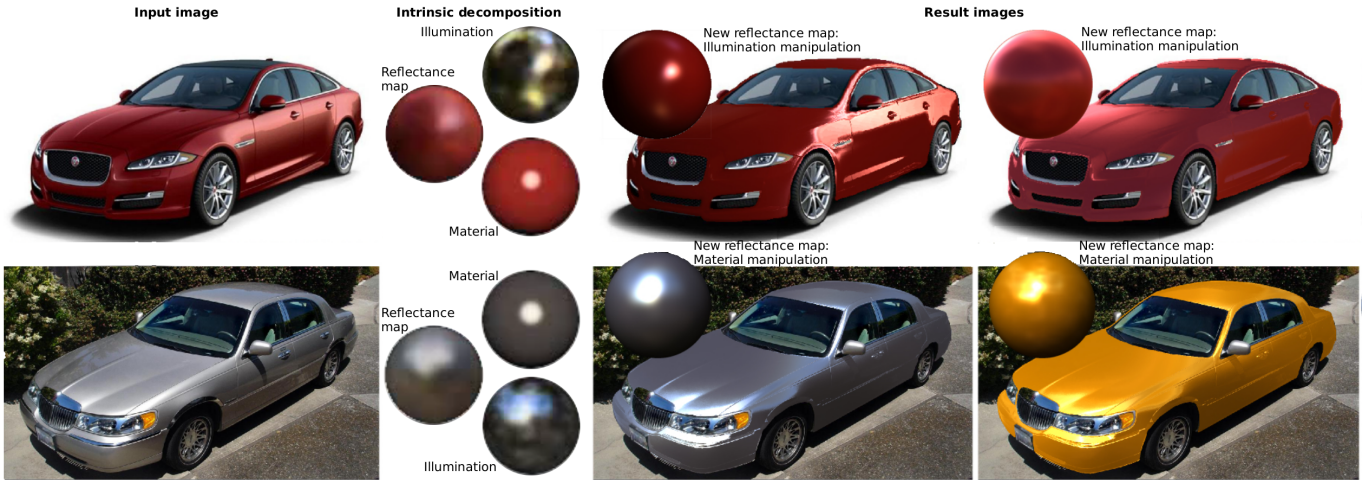


Figure 6.15: Illumination *(top row)* and material *(bottom row)* change *(3rd-4th columns)*, originally estimated from real photos *(1st column)*. For more details see *Manipulating material and illumination from real photos* in Sec. 6.7.3. The project's webpage contains a video with more such examples.

Chapter 7

Conclusion

Undoing the image formation process and essentially factoring images into their intrinsic components, *i.e.* 3D shape, surface reflectance and environmental illumination, is a relatively trivial task for humans. However, there are obstacles that should be overcome before the perception and reasoning skills of computers match those of humans. On the one hand, traditional approaches in literature have relied on simplifying assumptions to enable computers to perform this factorization, like using parametric BRDF models for reflectance and assuming remote point light sources for illumination. On the other hand, recent approaches with less strict reflectance and illumination assumptions still require the capture of HDR images and the use of dedicated hardware setups, rendering the scanning procedure rather impractical and inaccessible to non-expert users.

In this thesis, we focus on readily available capturing devices recording LDR images and we output refined geometry, non-parametric BRDFs and environmental illumination maps. This chapter summarizes our efforts to solve this inverse rendering problem, so that these individual intrinsic components if modified and re-synthesized result in a photo-realistic, yet faithful, rendering of the original scene. We start by presenting our contributions towards this goal. Then, we present the gained knowledge from this journey and all the lessons learned. We additionally describe in more detail the limitations of our proposed approaches and we provide the directions for future work.

7.1 Summary of Contributions

In this thesis, we address the problem of factoring an image into its intrinsic geometry, reflectance and illumination and we have explored the means to achieve this. As a result, the contributions of this thesis expand from extracting 3D shape from 2D images to inferring surface reflectance properties and estimating environmental illumination. In the following paragraphs we describe in more detail our contributions for each of these components.

7.1.1 Extracting 3D Shape from 2D Images

The first part of the thesis is dedicated to the extraction of 3D shape from 2D images. In Chapter 3, we showed how to solve a variant of this problem: given a rough initial 3D shape of an object, obtained using SfM, we refined its geometry using principles from PS, such that the final outcome closely resembles the ground truth object both in terms of geometric and radiometric details. Our experiments suggest that, for this particular task, using reflectance cues, like specular reflections, poses strong constraints on the direction of the photometric normals, which in turn guides the movement of the 3D points closer to their ground truth positions, improving the overall geometry.

In addition, in Chapter 6 we indirectly examined the 3D shape extraction problem. The proposed *Indirect* approach calculates per pixel normals as a by-product of the reflectance map estimation step when a single image depicting an object from a known class (*e.g.* cars) is given as input. It has to be noted, however, that our goal in this case was not normals, and as a result 3D shape, but rather reflectance and illumination estimation.

7.1.2 Measuring Surface Reflectance Properties

A significant amount of work presented in this thesis, in particular Chapters 3, 4 and 6, address the problem of measuring surface reflectance properties. The latter comes in the form of a 1D, 2D or 3D parametric or non-parametric BRDF. In Chapter 3, we presented an approach that samples and refines non-parametric BRDF slices and iteratively uses the estimates to refine the photometric normals and 3D points positions. The estimated BRDF slices are also used to create faithful renderings of the scanned object mimicking the illumination of the camera's flash during the scanning procedure. Starting from these BRDF slices, in Chapter 4, we introduced a GPLVM to infer the missing part of the higher dimensional non-parametric 2D or 3D BRDFs. By doing so, we showed

that we are able to relight the scanned object from different viewing/lighting configurations than the camera/flash one used for scanning and even reliably render it under natural illumination.

Moreover, in Chapter 6 we presented how to estimate a parametric Phong BRDF from a single LDR image depicting an object from a known class. To achieve this, we proposed a two-step approach, where we first estimate the object's reflectance map, and then further decompose the latter into reflectance and illumination. We demonstrated the effectiveness of our approach in estimating parametric Phong BRDFs through various applications, *e.g.* interactive material manipulation of car models from real photos.

7.1.3 Capturing Environmental Illumination

Chapters 5 and 6 cover the task of capturing environmental illumination when the input is a single LDR image. In Chapter 6, we explained how starting from a single LDR image depicting a single-material object from a known class we can in a first step estimate the object's reflectance map, and in a second step decompose the reflectance map into parametric reflectance and natural HDR illumination. Later in Chapter 5, we moved one step further and used everyday objects - i.e. far-from-perfect-mirrors both in terms of shape and materials - to act as light probes. We proposed a deep CNN that combines prior knowledge about the statistics of illumination and reflectance with an input that makes explicit use of the two key observations: (i) images rarely show a single material, but rather multiple ones that all reflect the same illumination, (ii) parts of the illumination are often directly observed in the background, without being affected by reflection. We showed how both the multi-material composition of the surfaces and using a background are essential to improve illumination estimations.

7.2 Observations

In this section we discuss the general observations regarding the methods proposed in the thesis. We present in more detail the evolution of the methods that occurred during the timeline of the thesis and what are the similarities and differences between them. The purpose of this section is to give more clarity for the research choices that we made and how they evolved throughout the thesis.

Surface reflectance under different lighting conditions Our first method to estimate surface reflectance was presented in Chapter 3, with its

extension following in Chapter 4 whereas our second method was introduced in Chapter 6. The main difference between the two lines of work is the lighting conditions under which reflectance is estimated. In the initial approach, the object is illuminated by the camera's flash which is considered to be dominant over any other illumination in the scene. In the second approach, the object is placed in a more natural scene where light is coming from every direction. Therefore, instead of estimating surface reflectance inside a "darkroom" under the point lighting assumption of the camera's flash and consequently infer the missing measurements using statistical models, we move "in-the-wild" and try to recover the reflective properties of a surface under natural environmental lighting, *e.g.* in an outdoors scene.

Both methods have advantages and disadvantages with respect to each other. The first approach is limited to controlled lighting environments where every illumination from the scene should be minimal compared to the camera's flash, more closely resembling the traditional gonioreflectometer devices used to scan reflectance. At the same time, however, it allows for the estimation of a more accurate non-parametric BRDF slice that expresses a wider range of materials and can also be used to compute photometric normals and optimize the geometry of the scanned object. Since the BRDF slice is a low dimensional BRDF, the missing higher dimensional reflectance properties have to be hallucinated, *e.g.* infer the missing measurements using statistical models. The second approach works in uncontrolled lighting environments, such as an outdoors natural scene, which is far more practical and less restricting compared to the first approach. Although, due to the highly under-constrained nature of this problem (*e.g.* a surface appears green because the material is green and the illumination is white or vice versa), some limiting assumptions have to be made in this case. For example, the object class should be known a priori and the surface reflectance is approximated by a parametric Phong model which imposes restrictions on the space of materials but ensures the plausibility of the final result.

Environmental illumination from a single LDR image In Chapter 6, we presented our first method for estimating environmental illumination from a single LDR image. Because we do not assume one or more components (shape, reflectance or illumination) to be known, to constrain the space of possible solutions we focused on input images that depict a single-material object of a given class with a specular material and under natural illumination. This allowed us to adopt a two-step approach, where we first estimate the object's reflectance map, and then further decompose the latter into reflectance and the desired illumination. Later, however, we realized that these simplifying assumptions that helped us in controlling this highly under constrained decomposition problem were quite limiting. For example, assuming a known object class limits the

method's applicability to a wider range of classes - most objects are not made of a single material and do not necessarily exhibit highly reflective behavior, the sampled reflectance maps are rarely densely filled, *etc.*

For the reasons described above, we decided to go beyond these simplifying assumptions and try to estimate environmental illumination from a single LDR image in a more realistic setting. For this task, we exploit two properties often found in everyday images. First, images rarely show a single material, but rather multiple ones that all reflect the same illumination. In fact, the appearance of each material can range from diffuse till specular and is observed only for some surface orientations, not all. The latter results in sparsely filled reflectance maps. Second, parts of the illumination are often directly observed in the background, without being affected by reflection. Typically, this directly observed part of the illumination is even smaller, but unlike before, we do not throw away this useful part of information that is captured in our input images anyway. In Chapter 5, we presented our second method that combines prior knowledge about the statistics of illumination and reflectance with an input that makes explicit use of these two key observations. Our results indicate that both the multi-material composition and using the background are essential to improve illumination estimations. Moreover, we observe that our method does not only retain efficiency across an increasing number of materials, but in fact uses the mutual information to produce even an increase in quality, which is in agreement with observations that humans are better in factoring illumination, shape and reflectance from complex aggregates than for simple ones [150].

Another observation to keep in mind, however, is that for the second approach to work we have to manually align a known 3D model of the object with the image, which is not the case for the first approach.

7.3 Lessons Learned

During the implementation of this thesis we addressed several problems on how to undo the image formation process and essentially factor images into their intrinsic components, *i.e.* 3D shape, surface reflectance and environmental illumination. This process taught us several lessons[1] regarding clarity, completeness and reproducibility that are worth sharing. The current section is to be taken as a guideline for future works that are planning to integrate the ideas and methods presented in this thesis.

[1]Except from the fact that doing a PhD is a hard job!

Reflectance cues pose strong photometric constraints Methods based on multi-view PS and using dedicated setups consisting of multiple lights have proven to generate accurate results for both diffuse and specular surfaces [43, 64, 169, 111]. For specular materials they rely on the assumption that the surface still exhibits an approximately Lambertian behaviour for at least a subset of the viewing/lighting combinations. However, when it comes to setups with a single light, like our camera/flash configuration (Chapter 3), finding such a subset of the viewing/lighting combinations where the scanned object exhibits an approximately Lambertian behaviour becomes challenging. In this particular case, we have found that specular reflections pose strong constraints on the direction of the photometric normals. As such, approaches that naturally handle specularities (Chapter 3), instead of discarding them as outliers [104, 43], can leverage them to arrive at better photometric normals estimates. The latter then guide the movement of the 3D points closer to their ground truth positions, improving the overall geometry (see Sec. 3.7.1).

When refining multiple components (shape, reflectance, illumination), the optimization should proceed in discrete steps In Sec. 3.6, we presented our reflectance and geometry refinement technique that involves the optimization of the base material (1D) BRDFs, photometric normals, material weights and 3D points positions, such that the estimated appearance for each point in each image fits the input observations. Through our experimentation, we have observed that optimizing multiple parameters at once is inefficient for two reasons: (1) the system is unstable, getting more easily stuck at local minima, resulting in implausible outcomes, and (2) a full global refinement is computationally very expensive (usually orders of magnitude higher in computation time). From our experience, in such multi-parameter optimization problems it is important to optimize each class of parameters independently and constraint the space of possible solutions in order to arrive at plausible results. The same principles apply for Chapter 6. There, we opted for a two-step approach where we first estimate a reflectance map from a 2D image (see Sec. 6.4) and second we decompose the reflectance map into reflectance and illumination (see Sec. 6.5). This separation of tasks helps in keeping the training process stable in our experiments.

Organizing materials into classes helps reflectance inference In Chapter 4, we showed that it is possible to predict the missing part of higher dimensional (2D or 3D) BRDFs starting from a single (1D) BRDF slice. To do so, we relied on a GPLVM (see Sec. 4.3.2), where BRDFs of different dimensionality (1D, 2D, 3D) can be regressed to a shared manifold. When designing our model, however, we found out that clustering the BRDFs into classes of similar material behaviour (e.g. plastics, paints, synthetic and natural fibers) allows us

to leverage the unique reflectance properties of each class of materials. Inspired by this observation, we opted for a discriminative prior that encourages the latent positions of the examples of the same class (e.g. plastics) to be close and those of different classes (e.g. plastics and paints) to be far on the shared manifold (see Sec. 4.3.3). The importance of using a discriminative prior for reflectance inference is reflected in the results (see Sec. 4.4), where our approach outperforms methods that do not separate the different classes [5, 21, 109].

Multiple materials and background information are essential for estimating illumination Factoring the environmental illumination from an image has received renewed interest. Recent approaches [69, 93] allow for the estimation of one component (*i.e.* shape, reflectance, illumination) if at least one other component is known and remains the same across the image (typically the shape). Although they show promising results, these methods are bound by strong constraints. The object should consist of a single material and be segmented from the background. Moreover, the information that the background provides is thrown away although it is captured in the input images anyway. Based on our experience, however, and supported by the findings from Chapter 5, having multiple materials as well as using the background information are essential for estimating illumination. In Sec. 5.5, we showed how to fuse the information from multiple materials and background together with reflectance and illumination priors in order to arrive at better illumination estimates. Our results (see Sec. 5.6) indicate how both having multiple materials and using a background help to improve the illumination estimations. In fact, our method leverages the mutual information across an increasing number of materials to produce even an increase in the estimated illumination's quality, the same way humans are better in factoring illumination, shape and reflectance from complex aggregates than for simple ones [150].

The single image decomposition into shape, reflectance, illumination requires strong assumptions In Chapter 6, we presented our approach for estimating shape (indirectly), reflectance and illumination from a single image, essentially dealing with the inverse rendering problem as a whole. It is obvious that estimating this amount of information from a single image is challenging as the same visual result might be due to many different combinations of the individual components. As such, for these highly under-constrained problems one has to impose strong assumptions to arrive at meaningful and plausible results. In our case, this translates to assumptions about the object (it should come from a predefined class and consist of a single material), the image (a segmentation from the background has to be given as input), the camera (the object is seen under orthographic projection from an infinitely far-away observer),

the reflectance model (the incoming light only depends on direction), *etc.* We have found these simplifications necessary for a successful decomposition.

7.4 Research Questions Revisited

At the beginning of this thesis we set the objective of investigating methods for extracting 3D shape, inferring surface reflectance properties, and estimating environmental illumination from a single or a set of images. To this end, we formulated four research questions that were presented in the introduction. From then on, we went on a journey whose aim was to provide the answers to these questions. Based on these answers, we would like to revisit the research questions and address them from a new perspective.

1. Can we extract 3D shape and surface reflectance from a small set of uncalibrated images and under which lighting conditions?

This research question was explored under two additional constraints. First, only readily available consumer equipment has to be used for the scanning. Second, both 3D shape and surface reflectance have to be recovered. Despite these extra constraints, the answer is still positive. In Chapter 3, we introduced a method that uses just a DSLR camera or smartphone and the illumination of their flash to capture an object's 3D shape and reflectance characteristics at the same time. Starting from a low-resolution mesh generated from SfM and MvS, we applied a new PS-based technique to refine both geometry and reflectance. We experimentally validated our approach by modeling several challenging examples, both synthetic and real, ranging from diffuse till highly specular surfaces. Note that acquiring accurate 3D shape and photo-realistic reflectance with this setup is a hard problem. Nevertheless, our method performs better than existing approaches designed for complicated hardware setups.

2. To what extent can we infer high-dimensional reflectance information from a single image?

In Chapter 4, we showed how to infer the missing 2D and 3D part of a BRDF starting from a single (1D) BRDF slice. To tackle this problem we proposed a GPLVM to infer the higher dimensional properties of the material's BRDF, based on the statistical distribution of known material characteristics observed in real-life samples. We also used a discriminative prior that leverages the unique reflectance properties of each class of materials. Although inferring

higher dimensional BRDFs from such modest training is not a trivial problem, our method performs better than existing parametric, semi-parametric and non-parametric approaches. We also presented interesting applications of our method on material relighting and flash-based photography.

3. How can we estimate the environmental illumination of a multi-material object given an image as the sole input?

The answer to this research question comes from two properties often found in everyday images. First, images usually show objects consisting of multiple materials that all reflect the same illumination, and second, a small part of the illumination is directly observed in the object's background. In Chapter 5, we proposed a deep learning approach that incorporates information from these two sources (*i.e.* multiple materials and background) as well as from reflectance and illumination priors to arrive at a HDR environment map from a single LDR image. Our qualitative and quantitative results showed how both multi-material and using a background are essential to improve illumination estimations. The presented method enables everyday objects - *i.e.* far-from-perfect-mirrors both in terms of shape and materials - to act as light probes.

4. Is it possible to decompose a single image into its intrinsic 3D shape, surface reflectance and environmental illumination and if so, what assumptions should be made to make this decomposition feasible?

Under the assumptions of (i) an object consisting of a single material, (ii) coming from a known class, (iii) which is segmented from its background, (iv) and seen under orthographic projection from an infinitely far-away observer, (v) where the incoming light only depends on direction, it is possible to recover all three components (shape, reflectance, illumination) from a single image. In Chapter 6, we propose a two-step deep learning approach for this highly under-constrained problem, where we first estimate the object's reflectance map, and then further decompose the latter into reflectance and illumination. We demonstrated the effectiveness of our approach for both steps by extensive qualitative and quantitative evaluation in both synthetic and real data as well as through numerous applications, that show improvements over existing approaches. Besides allowing for a better understanding and analysis of 2D imagery, the ability to estimate reflectance maps lends itself to a broad spectrum of applications, including appearance transfer, inpainting and augmented reality, while its further decomposition into reflectance and illumination enables powerful image editing applications, such as material transfer and illumination editing.

7.5 Limitations

In the previous chapters, we presented the core of this thesis and our approaches addressing the inverse rendering problem (*i.e.* factor images into 3D shape, surface reflectance and environmental illumination) and its related sub-problems (*i.e.* extract 3D shape from 2D images, infer surface reflectance properties, estimate environmental illumination). We showed successful applications of the proposed methods with a large number of qualitative and quantitative results. However, as can be expected, there is room for further improvement on these methods. In this section, we highlight the main weak points of our work to give the reader a more complete view over the thesis.

Working inside the "darkroom" In this thesis, we presented an approach to first capture an object's 3D shape and BRDF slices in Chapter 3 and consequently infer the missing higher dimensional BRDF properties in Chapter 4. In both cases, the input is either a single image or a sequence of images taken under the illumination of the camera's flash. This implies the scanning inside a "darkroom" where the illumination of the flash is dominant over any other illumination in the scene. Although more practical than existing approaches this camera/flash setup is still restricting in everyday life as most pictures are taken in naturally illuminated indoors or outdoors scenes. Assuming that the existing approaches can be applied in such lighting environments would be an overstatement although there is always the option of adjusting the camera's settings (*i.e.* ISO, exposure) so that the flash appears as the dominant source of illumination even in this case. Ideally, however, one should consider designing approaches that work with minimal equipment - just a camera - under natural illumination. An added advantage would be the use of videos instead of images as the former is far more practical for the user (*e.g.* recording around fifty to a hundred images for such approaches to work is undeniably impractical compared to a video of a few seconds depicting the object from different viewpoints).

Data quantity and quality Perhaps the strongest limitations of this thesis stem from the quantity and quality of the used datasets. Unfortunately, large databases of BRDF and HDR illumination samples are by and large lacking still. The existing solutions include MERL BRDF database that has 100 material samples from 4 classes that are all related to cars (*e.g.* car paints, metals for the exterior parts and fabrics for the interior parts) and approximately 20 HDR environment maps, courtesy of Debevec [28], that show some indoors and outdoors scenes. Regarding reflectance, we had to limit ourselves to the 100 MERL samples since scanning more would require a gonioreflectometer that is not available in our lab, as in the vast majority of research labs worldwide too.

To cope with the limited number of reflectance samples, for our experiments in Chapter 4 we had to perform a 5-fold cross-validation where 60 samples (out of the 100 MERL materials) are used for training, 20 for validation and 20 for testing. Although this limited training set was enough for generating results than outperform existing parametric, semi-parametric and non-parametric approaches, a more thorough evaluation would require at least an order of magnitude more material samples which would allow us to reach stronger conclusions. On top of this, since the recorded MERL samples are at most 3D BRDFs we were not able to test our method's ability to infer even higher dimensional BRDFs. The exact same reflectance limitations apply to Chapter 6 where 67 MERL samples were used for training and 33 for testing. Regarding illumination, working with existing HDR environment maps was not an option, especially for training our CNN models that require many input images. To overcome this problem, in Chapter 6 we collected 70 free HDR environment maps in total - 60 for training and 10 for testing - from the commercial content supplier HDR Maps (`https://www.hdrmaps.com/`) and in Chapter 5 we further enhanced this set by collecting more HDR environment maps, resulting in a total of 105 publicly available HDR environment maps. The main problem in this case is that most of these environment maps come from professional or amateur photographers that were not interested in capturing all the available dynamic range but rather limited themselves to 7 f-stops at most. As such, although we work with HDR data their quality is not the best possible (*e.g.* for proper relighting applications maybe more than 7 f-stops are needed). Furthermore, our learning based approaches would greatly benefit from the added dynamic range if the latter was available. Finally, it should be noted that for a few environment maps the recorded dynamic range is 3 or 5 f-stops which creates some inconsistency between the data. The only alternative would be to manually scan a much larger database of HDR environment maps with consistent recordings and the maximum available dynamic range but the latter requires a lot of manual work.

Methods' assumptions The methods presented in this thesis to tackle the problem of undoing the image formation process and essentially factor images into their intrinsic components, *i.e.* 3D shape, surface reflectance and environmental illumination, rely on rather strong assumptions. As explained in the previous sections, on the one hand this is necessary to keep the complexity of the proposed approaches under control and arrive at plausible decomposition results, but on the other hand it would be useful to see to what extent we can relax some of these assumptions and still arrive at the desired results. Below we give some indicative examples. In Chapters 3 and 4, we assume the light is only coming from the flash. In Chapters 5 and 6, we assume orthographic cameras and do not explicitly handle the indirect illumination, such as light sources that

are shadowed by occluders and rays that bounce multiple times around a scene while making their trip from a light source to the camera. Also, in both cases we are given as input a segmentation mask separating the different materials or the background from the object of interest. In Chapter 6, we assume the object class is known in advance. In general, it would be interesting to take most of these aspects into consideration and design more general approaches.

7.6 Future Work

In the last section of this thesis, we draw possible directions for future work based on the contributions, the observations and the limitations of the presented work and we propose several research paths that can be followed.

The first direction for future work lies in designing a method that can estimate 3D shape and surface reflectance from a sequence of images depicting an object under natural illumination and without relying on the illumination of the camera's flash. This is in essence an extension of the approach presented in Chapter 3 and would require a new optimization technique that takes into account lighting from multiple directions to refine the base materials BRDFs and weights, photometric normals and 3D points positions. Especially for the latter, instead of moving the 3D points after estimating new photometric normals it would be preferable to sequentially optimize them on-the-fly together with the other parameters and not as a standalone final step. In this way, we can also relax the decent geometry initializations required by the current technique. To further enhance the practicality of the method, a great addition would be the use of videos as input instead of images. This would greatly reduce the acquisition effort on the user side.

The second direction for future work is related to the inference of higher dimensional BRDFs from low dimensional inputs. In the current method, presented in Chapter 4, the proposed GPLVM is trained from a limited number of MERL samples. As explained in the previous section, this is mainly due to the lack of large scale BRDF databases. As such, scanning our own BRDF database with an increased number of material samples from more classes compared to MERL would be a great step for further investigating the reflectance inference problem. Moreover, the latter would allow us to leverage the unique reflective properties of each class of materials for other problems too, like material classification from reflectance cues, that show great potential. Both gaussian processes and deep learning could be used towards this goal.

The third direction, addresses the limitation of the available dynamic range in the environment maps used in Chapters 5 and 6. A straightforward solution, in this

case too, would be to manually scan a much larger database of HDR environment maps with consistent recordings and the maximum available dynamic range. Despite the load of manual work required for acquiring these data, our learning based approaches would greatly benefit from the added dynamic range. What would be needed, however, is the design of a new loss in our CNN models that takes into account the added dynamic range and nicely balances between learning the later and the color distribution of the environment map. If possible, a useful addition would be the inclusion of many real examples with annotated material and background/foreground segmentations that could be used to train our CNN models to also learn these segmentations instead of inputting them.

Finally, as presented in Chapter 6, there is strong potential on recovering surface reflectance and HDR environmental illumination from a single LDR image when the 3D shape is not known but instead the object's class is known. However, the proposed pipeline works in two discrete steps first estimating a reflectance map and second decomposing the reflectance map into Phong BRDF parameters and a HDR environment map. What could be done instead, is to design an alternative approach that is trained end-to-end and takes into account the estimates of each iteration step to iteratively refine the outputs. This could potentially increase the efficiency of the method and lead to overall better results for this hard decomposition problem. Also, as mentioned in the previous section, the approach would benefit from dropping the assumption of a known object class. This way we could extend to different object classes and not only work with *e.g.* cars.

Bibliography

[1] Agarwal, S., Mierle, K., et al. Ceres solver, 2013. 50

[2] Agrawal, A., Raskar, R., Nayar, S. K., and Li, Y. Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Transactions on Graphics (TOG) 24*, 3 (2005), 828–835. 35

[3] Aliaga, D. G., and Xu, Y. Photogeometric structured light: A self-calibrating and multi-viewpoint framework for accurate 3d modeling. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8. 36

[4] Alldrin, N., Zickler, T., and Kriegman, D. Photometric stereo with non-parametric and spatially-varying reflectance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8. 36, 38

[5] Ashikhmin, M., and Premoze, S. Distribution-based brdfs. *Unpublished Technical Report, University of Utah 2* (2007), 6. 76, 77, 78, 79, 80, 143

[6] Ashikhmin, M., and Shirley, P. An anisotropic phong brdf model. *Journal of graphics tools 5*, 2 (2000), 25–32. 23, 67

[7] Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* (2015). 28

[8] Barron, J. T., and Malik, J. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 17–24. 3, 87

[9] Barron, J. T., and Malik, J. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence 37*, 8 (2015), 1670–1687. 86, 109

[10] Barrow, H., and Tenenbaum, J. Computer vision systems. *Computer vision systems 2* (1978). 85, 86, 104, 107, 108

[11] Bay, H., Tuytelaars, T., and Van Gool, L. Surf: Speeded up robust features. In *European conference on computer vision* (2006), Springer, pp. 404–417. 18

[12] Bell, S., Bala, K., and Snavely, N. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 159. 104, 107

[13] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG) 32*, 4 (2013), 111. 85

[14] Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014. 73

[15] Blinn, J. F. Models of light reflection for computer synthesized pictures. In *ACM SIGGRAPH Computer Graphics* (1977), vol. 11, ACM, pp. 192–198. 23, 43, 67, 76

[16] Blinn, J. F., and Newell, M. E. Texture and reflection in computer generated images. *Communications of the ACM 19*, 10 (1976), 542–547. 25

[17] Boivin, S., and Gagalowicz, A. Image-based rendering of diffuse, specular and glossy surfaces from a single image. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), ACM, pp. 107–116. 67

[18] Chan, D., Buisman, H., Theobalt, C., and Thrun, S. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008* (2008). 115

[19] Chandraker, M., Agarwal, S., and Kriegman, D. Shadowcuts: Photometric stereo with shadows. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8. 34

[20] Chandraker, M., Bai, J., and Ramamoorthi, R. A theory of differential photometric stereo for unknown isotropic brdfs. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 2505–2512. 36, 38

[21] Chandraker, M., and Ramamoorthi, R. What an image reveals about material reflectance. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 1076–1083. 67, 68, 70, 76, 77, 78, 79, 80, 143

[22] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). 29, 90, 121

[23] Chen, Q., and Koltun, V. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 241–248. 95

[24] Chen, T., Goesele, M., and Seidel, H.-P. Mesostructure from specularity. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, IEEE, pp. 1825–1832. 38, 55

[25] Cook, R. L., and Torrance, K. E. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG) 1*, 1 (1982), 7–24. 23, 67, 76

[26] Cryer, J. E., Tsai, P.-S., and Shah, M. Integration of shape from shading and stereo. *Pattern recognition 28*, 7 (1995), 1033–1043. 34

[27] Dana, K. J., Van Ginneken, B., Nayar, S. K., and Koenderink, J. J. Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG) 18*, 1 (1999), 1–34. 7, 35, 85, 107, 110

[28] Debevec, P. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008 classes* (2008), ACM, p. 32. 8, 146

[29] Debevec, P., Yu, Y., and Borshukov, G. Efficient view-dependent image-based rendering with projective texture-mapping. In *Rendering Techniques' 98*. Springer, 1998, pp. 105–116. 25, 85, 86, 107, 110

[30] Debevec, P. E., Taylor, C. J., and Malik, J. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (1996), ACM, pp. 11–20. 9, 84, 90, 111, 117

[31] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 248–255. 26

[32] Dong, C., Loy, C. C., He, K., and Tang, X. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision* (2014), Springer, pp. 184–199. 87

[33] Dong, Y., Wang, J., Tong, X., Snyder, J., Lan, Y., Ben-Ezra, M., and Guo, B. Manifold bootstrapping for svbrdf capture. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 98. 35

[34] Dosovitskiy, A., Tobias Springenberg, J., and Brox, T. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1538–1546. 109

[35] Dror, R. O., Leung, T. K., Adelson, E. H., and Willsky, A. S. Statistics of real-world illumination. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference On* (2001), vol. 2, IEEE, pp. II–II. 85, 86, 87, 107, 117

[36] Dror, R. O., Willsky, A. S., and Adelson, E. H. Statistical characterization of real-world illumination. *Journal of Vision 4*, 9 (2004), 11–11. 67

[37] Eigen, D., and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 2650–2658. 85, 87, 88, 105, 110, 114, 117

[38] Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems* (2014), pp. 2366–2374. 87, 105, 110, 117

[39] Eisemann, E., and Durand, F. Flash photography enhancement via intrinsic relighting. *ACM transactions on graphics (TOG) 23*, 3 (2004), 673–678. 35

[40] Ek, C. H., and Lawrence, P. *Shared Gaussian process latent variable models*. PhD thesis, PhD thesis, 2009. 70

[41] Eleftheriadis, S., Rudovic, O., and Pantic, M. View-constrained latent variable model for multi-view facial expression classification. In *International Symposium on Visual Computing* (2014), Springer, pp. 292–303. 72

[42] Eleftheriadis, S., Rudovic, O., and Pantic, M. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing 24*, 1 (2015), 189–204. 70, 72

[43] Esteban, C. H., Vogiatzis, G., and Cipolla, R. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 3 (2008), 548–554. 21, 34, 38, 39, 46, 51, 52, 53, 54, 55, 142

[44] Fairchild, M. D. *Color appearance models*. John Wiley & Sons, 2013. 75

[45] Filip, J., Vavra, R., Haindl, M., Zid, P., Krupika, M., and Havran, V. Brdf slices: Accurate adaptive anisotropic appearance acquisition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1468–1473. 35

[46] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6 (1981), 381–395. 19

[47] Fitzgibbon, A. W., and Zisserman, A. Automatic camera recovery for closed or open image sequences. In *European conference on computer vision* (1998), Springer, pp. 311–326. 39

[48] Fleming, R. W., Dror, R. O., and Adelson, E. H. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision 3*, 5 (2003), 3–3. 67

[49] Furukawa, Y., and Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence 32*, 8 (2010), 1362–1376. 34

[50] Georgoulis, S., Proesmans, M., and Van Gool, L. Tackling shapes and brdfs head-on. In *3D Vision (3DV), 2014 2nd International Conference on* (2014), vol. 1, IEEE, pp. 267–274. 78

[51] Georgoulis, S., Vanweddingen, V., Proesmans, M., and Van Gool, L. A gaussian process latent variable model for brdf inference. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3559–3567. 104

[52] Ghosh, A., Chen, T., Peers, P., Wilson, C. A., and Debevec, P. Estimating specular roughness and anisotropy from second order spherical

gradient illumination. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 1161–1170. 33, 36

[53] Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 580–587. 109

[54] Goldman, D. B., Curless, B., Hertzmann, A., and Seitz, S. M. Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 6 (2010), 1060–1071. 34, 35, 36

[55] Greene, N. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications 6*, 11 (1986), 21–29. 25

[56] Haber, T., Fuchs, C., Bekaer, P., Seidel, H.-P., Goesele, M., and Lensch, H. P. Relighting objects from image collections. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 627–634. 86, 108

[57] Hara, K., Nishino, K., and Ikeuchi, K. Mixture of spherical distributions for single-view relighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 1 (2008), 25–35. 67

[58] Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 447–456. 109

[59] He, X. D., Torrance, K. E., Sillion, F. X., and Greenberg, D. P. A comprehensive physical model for light reflection. In *ACM SIGGRAPH computer graphics* (1991), vol. 25, ACM, pp. 175–186. 23, 67, 76

[60] Hertzmann, A., and Seitz, S. M. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 8 (2005), 1254–1264. 34, 35, 108

[61] Higo, T., Matsushita, Y., and Ikeuchi, K. Consensus photometric stereo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 1157–1164. 34

[62] Higo, T., Matsushita, Y., Joshi, N., and Ikeuchi, K. A hand-held photometric stereo camera for 3-d modeling. In *Computer Vision, 2009*

IEEE 12th International Conference on (2009), IEEE, pp. 1234–1241. 36, 37, 38, 52

[63] Holroyd, M., Lawrence, J., Humphreys, G., and Zickler, T. A photometric approach for estimating normals and tangents. *ACM Transactions on Graphics (TOG) 27*, 5 (2008), 133. 36

[64] Holroyd, M., Lawrence, J., and Zickler, T. A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 99. 32, 33, 36, 66, 142

[65] Hoppe, H. New quadric metric for simplifiying meshes with appearance attributes. In *Proceedings of the conference on Visualization'99: celebrating ten years* (1999), IEEE Computer Society Press, pp. 59–66. 52

[66] Horn, B. K. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 2, 3, 21

[67] Horn, B. K., and Sjoberg, R. W. Calculating the reflectance map. *Applied optics 18*, 11 (1979), 1770–1779. 85, 86, 88, 104, 105, 108, 110

[68] Jin, H., Cremers, D., Yezzi, A. J., and Soatto, S. Shedding light on stereoscopic segmentation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on* (2004), vol. 1, IEEE, pp. I–I. 46

[69] Johnson, M. K., and Adelson, E. H. Shape estimation in natural illumination. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 2553–2560. 86, 108, 143

[70] Kajiya, J. T. The rendering equation. In *ACM Siggraph Computer Graphics* (1986), vol. 20, ACM, pp. 143–150. 111

[71] Karsch, K., Sunkavalli, K., Hadap, S., Carr, N., Jin, H., Fonte, R., Sittig, M., and Forsyth, D. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG) 33*, 3 (2014), 32. 87, 95

[72] Kazhdan, M., and Hoppe, H. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG) 32*, 3 (2013), 29. 39

[73] Khan, E. A., Reinhard, E., Fleming, R. W., and Bülthoff, H. H. Image-based material editing. *ACM Transactions on Graphics (TOG) 25*, 3 (2006), 654–663. 86, 108

[74] Koenderink, J. J., and Van Doorn, A. J. Phenomenological description of bidirectional surface reflection. *JOSA A 15*, 11 (1998), 2903–2912. 7, 32, 66

[75] Kopf, J., Cohen, M. F., Lischinski, D., and Uyttendaele, M. Joint bilateral upsampling. In *ACM Transactions on Graphics (ToG)* (2007), vol. 26, ACM, p. 96. 115

[76] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105. 26, 109

[77] Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems* (2015), pp. 2539–2547. 110

[78] Lafortune, E. P., Foo, S.-C., Torrance, K. E., and Greenberg, D. P. Non-linear approximation of reflectance functions. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), ACM Press/Addison-Wesley Publishing Co., pp. 117–126. 23, 67, 76

[79] Lalonde, J.-F., Efros, A. A., and Narasimhan, S. G. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision 98*, 2 (2012), 123–145. 87

[80] Lanman, D., Sibley, P. G., Crispell, D., Zhao, Y., and Taubin, G. Multi-flash 3d photography: Capturing shape and appearance. In *ACM SIGGRAPH 2006 Research posters* (2006), ACM, p. 99. 35

[81] Lawrence, J., Ben-Artzi, A., DeCoro, C., Matusik, W., Pfister, H., Ramamoorthi, R., and Rusinkiewicz, S. Inverse shade trees for non-parametric material representation and editing. In *ACM Transactions on Graphics (TOG)* (2006), vol. 25, ACM, pp. 735–745. 68

[82] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (2009), ACM, pp. 609–616. 109

[83] Lensch, H., Kautz, J., Goesele, M., Heidrich, W., and Seidel, H.-P. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics (TOG) 22*, 2 (2003), 234–257. 35, 38, 43, 44, 86

[84] Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., et al. The digital michelangelo project: 3d scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), ACM Press/Addison-Wesley Publishing Co., pp. 131–144. 35

[85] Lhuillier, M., and Quan, L. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence 27*, 3 (2005), 418–433. 34

[86] Li, B., Shen, C., Dai, Y., van den Hengel, A., and He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1119–1127. 85, 87, 88, 110, 114, 117

[87] Lim, S. Characterization of noise in digital photographs for image processing. In *Electronic Imaging 2006* (2006), International Society for Optics and Photonics, pp. 60690O–60690O. 59

[88] Liu, C., Freeman, W. T., Szeliski, R., and Kang, S. B. Noise estimation from a single image. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 1, IEEE, pp. 901–908. 59

[89] Liu, F., Shen, C., and Lin, G. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 5162–5170. 87, 110, 117

[90] Lombardi, S., and Nishino, K. Reflectance and natural illumination from a single image. In *European Conference on Computer Vision* (2012), Springer, pp. 582–595. 56, 124, 127, 128, 129, 130, 131

[91] Lombardi, S., and Nishino, K. Single image multimaterial estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (2012), IEEE, pp. 238–245. 35, 66, 67, 86, 127

[92] Lombardi, S., and Nishino, K. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *3D Vision (3DV), 2016 Fourth International Conference on* (2016), IEEE, pp. 305–313. 87, 117

[93] Lombardi, S., and Nishino, K. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*

38, 1 (2016), 129–141. 3, 56, 66, 67, 86, 87, 104, 105, 106, 108, 109, 117, 124, 127, 143

[94] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440. 109, 114

[95] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110. 18

[96] Lucas, B. D., Kanade, T., et al. An iterative image registration technique with an application to stereo vision. 19

[97] Ma, W.-C., Hawkins, T., Peers, P., Chabert, C.-F., Weiss, M., and Debevec, P. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques* (2007), Eurographics Association, pp. 183–194. 36

[98] Marschner, S. R., Westin, S. H., Lafortune, E. P., and Torrance, K. E. Image-based bidirectional reflectance distribution function measurement. *Applied Optics 39*, 16 (2000), 2592–2600. 32, 66

[99] Matusik, W. *A data-driven reflectance model*. PhD thesis, Citeseer, 2003. 24, 29, 30, 32, 35, 47, 66, 68, 75, 85, 86, 90, 107, 121, 122, 123, 127

[100] Melendez, F., Glencross, M., Ward, G. J., and Hubbold, R. J. High-resolution relightable buildings from photographs. In *ACM SIGGRAPH 2011 Talks* (2011), ACM, p. 39. 35

[101] Moons, T., Van Gool, L., Vergauwen, M., et al. 3d reconstruction from multiple images part 1: Principles. *Foundations and Trends® in Computer Graphics and Vision 4*, 4 (2010), 287–404. 18

[102] Narihira, T., Maire, M., and Yu, S. X. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 2992–2992. 87, 95, 107, 110

[103] Narihira, T., Maire, M., and Yu, S. X. Learning lightness from human judgement on relative reflectance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 2965–2973. 87, 110

[104] Nehab, D., Rusinkiewicz, S., Davis, J., and Ramamoorthi, R. Efficiently combining positions and normals for precise 3d geometry. In *ACM transactions on graphics (TOG)* (2005), vol. 24, ACM, pp. 536–543. 36, 38, 49, 51, 52, 53, 54, 55, 142

[105] Ngan, A., Durand, F., and Matusik, W. Experimental analysis of brdf models. *Rendering Techniques 2005*, 16th (2005), 2. 23, 67, 68

[106] Nicodemus, F. E. Directional reflectance and emissivity of an opaque surface. *Applied optics 4*, 7 (1965), 767–775. 6, 32, 35, 66

[107] Nishino, K. Directional statistics brdf model. In *Computer Vision, 2009 IEEE 12th International Conference on* (2009), IEEE, pp. 476–483. 23, 43, 67, 68, 76

[108] Nishino, K., Zhang, Z., and Ikeuchi, K. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2001), vol. 1, IEEE, pp. 599–606. 67

[109] Nöll, T., Köhler, J., and Stricker, D. Robust and accurate non-parametric estimation of reflectance using basis decomposition and correction functions. In *European Conference on Computer Vision* (2014), Springer, pp. 376–391. 66, 68, 76, 77, 78, 79, 80, 143

[110] Oxholm, G., and Nishino, K. Shape and reflectance estimation in the wild. *IEEE transactions on pattern analysis and machine intelligence 38*, 2 (2016), 376–389. 3, 32, 33, 36, 37, 39, 52

[111] Park, J., Sinha, S. N., Matsushita, Y., Tai, Y.-W., and So Kweon, I. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1161–1168. 32, 34, 38, 142

[112] Patel, V. M., Van Nguyen, H., and Vidal, R. Latent space sparse subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 225–232. 75

[113] Pathak, D., Krähenbühl, P., Yu, S. X., and Darrell, T. Constrained structured regression with convolutional neural networks. *arXiv preprint arXiv:1511.07497* (2015). 95

[114] Pellacini, F., Ferwerda, J. A., and Greenberg, D. P. Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the 27th annual conference on Computer graphics and*

interactive techniques (2000), ACM Press/Addison-Wesley Publishing Co., pp. 55–64. 67

[115] Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., and Toyama, K. Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG) 23*, 3 (2004), 664–672. 35

[116] Phong, B. T. Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (1975), 311–317. 23, 111

[117] Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R. Visual modeling with a hand-held camera. *International Journal of Computer Vision 59*, 3 (2004), 207–232. 6, 19

[118] Ramamoorthi, R., and Hanrahan, P. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), ACM, pp. 117–128. 87, 88, 105

[119] Rasmussen, C. E. Gaussian processes for machine learning. 50, 70, 73

[120] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J. Photographic tone reproduction for digital images. *ACM transactions on graphics (TOG) 21*, 3 (2002), 267–276. 121

[121] Rematas, K., Nguyen, C., Ritschel, T., Fritz, M., and Tuytelaars, T. Novel views of objects from a single image. *arXiv preprint arXiv:1602.00328* (2016). 108

[122] Rematas, K., Ritschel, T., Fritz, M., Gavves, E., and Tuytelaars, T. Deep reflectance maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4508–4516. 85, 86, 88, 95

[123] Rematas, K., Ritschel, T., Fritz, M., and Tuytelaars, T. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *Computer Vision and Pattern Recognition (CVPR)* (2014). 86, 108

[124] Ren, P., Wang, J., Snyder, J., Tong, X., and Guo, B. Pocket reflectometry. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 45. 35, 36, 37, 38, 52

[125] Richter, S. R., and Roth, S. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1128–1136. 86, 109

[126] Romeiro, F., and Zickler, T. Blind reflectometry. In *European conference on computer vision* (2010), Springer, pp. 45–58. 66, 68, 87, 105, 109

[127] Rusinkiewicz, S., Hall-Holt, O., and Levoy, M. Real-time 3d model acquisition. *ACM Transactions on Graphics (TOG) 21*, 3 (2002), 438–446. 35

[128] Rusinkiewicz, S. M. A new change of variables for efficient brdf representation. In *Rendering techniques' 98*. Springer, 1998, pp. 11–22. 41, 42

[129] Sato, I., Sato, Y., and Ikeuchi, K. Illumination from shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 3 (2003), 290–300. 87

[130] Schlick, C. An inexpensive brdf model for physically-based rendering. In *Computer graphics forum* (1994), vol. 13, Wiley Online Library, pp. 233–246. 23, 67, 76

[131] Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International Conference on Computational Learning Theory* (2001), Springer, pp. 416–426. 72

[132] Schwartz, C., Weinmann, M., Ruiters, R., and Klein, R. Integrated high-quality acquisition of geometry and appearance for cultural heritage. In *VAST* (2011), pp. 25–32. 62

[133] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 1, IEEE, pp. 519–528. 52, 53, 59

[134] Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., and Adelson, E. H. Image statistics for surface reflectance perception. *JOSA A 25*, 4 (2008), 846–865. 67

[135] Shi, B., Tan, P., Matsushita, Y., and Ikeuchi, K. Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In *European Conference on Computer Vision* (2012), Springer, pp. 455–468. 36

[136] Shon, A. P., Grochow, K., Hertzmann, A., and Rao, R. P. Learning shared latent structure for image synthesis and robotic imitation. *Advances in neural information processing systems 18* (2006), 1233. 70

[137] Sloan, P.-P. J., Martin, W., Gooch, A., and Gooch, B. The lit sphere: A model for capturing npr shading from art. In *Graphics interface* (2001), vol. 2001, pp. 143–150. 86, 108, 112

[138] Spencer, S. *ZBrush Character Creation: Advanced Digital Sculpting*. John Wiley & Sons, 2011. 86, 108

[139] Stark, M. M., Arvo, J., and Smits, B. Barycentric parameterizations for isotropic brdfs. *IEEE transactions on visualization and computer graphics 11*, 2 (2005), 126–138. 68

[140] Sun, J., Cao, W., Xu, Z., and Ponce, J. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 769–777. 87

[141] Sundararajan, S., and Keerthi, S. S. Predictive approaches for choosing hyperparameters in gaussian processes. *Neural computation 13*, 5 (2001), 1103–1118. 73

[142] Szeliski, R. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 15, 16, 22

[143] Tan, P., Quan, L., and Zickler, T. The geometry of reflectance symmetries. *IEEE transactions on pattern analysis and machine intelligence 33*, 12 (2011), 2506–2520. 36

[144] Tang, Y., Salakhutdinov, R., and Hinton, G. Deep lambertian networks. *arXiv preprint arXiv:1206.6445* (2012). 110

[145] Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., and Pollefeys, M. Live metric 3d reconstruction on mobile phones. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 65–72. 32

[146] Taubin, G. A signal processing approach to fair surface design. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* (1995), ACM, pp. 351–358. 52

[147] Tingdahl, D., Godau, C., and Van Gool, L. Base materials for photometric stereo. In *European Conference on Computer Vision* (2012), Springer, pp. 350–359. 35, 44

[148] Tingdahl, D., and Van Gool, L. A public system for image based 3d model generation. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications* (2011), Springer, pp. 262–273. 19, 39

[149] Van Meerbergen, G., Vergauwen, M., Pollefeys, M., and Van Gool, L. A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision 47*, 1-3 (2002), 275–285. 19

[150] Vangorp, P., Laurijssen, J., and Dutré, P. The influence of shape on the perception of material reflectance. In *ACM Transactions on Graphics (TOG)* (2007), vol. 26, ACM, p. 77. 86, 99, 141, 143

[151] Verbiest, F., and Van Gool, L. Photometric stereo with coherent outlier handling and confidence estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8. 34

[152] Vogiatzis, G., Hernandez, C., and Cipolla, R. Reconstruction in the round using photometric normals and silhouettes. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, IEEE, pp. 1847–1854. 46

[153] Wang, X., Fouhey, D., and Gupta, A. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 539–547. 85, 87, 88, 110, 114, 117

[154] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612. 95

[155] Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on* (2003), vol. 2, IEEE, pp. 1398–1402. 122, 129

[156] Ward, G. J. Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics 26*, 2 (1992), 265–272. 23, 67, 76

[157] Watt, A. *3D computer graphics*. Addison-Wesley Longman Publishing Co., Inc., 1993. 25

[158] Wills, J., Agarwal, S., Kriegman, D., and Belongie, S. Toward a perceptual space for gloss. *ACM Transactions on Graphics (TOG) 28*, 4 (2009), 103. 67

[159] Woodham, R. J. Photometric method for determining surface orientation from multiple images. *Optical engineering 19*, 1 (1980), 191139–191139. 6, 20, 34

[160] Wu, C., Wilburn, B., Matsushita, Y., and Theobalt, C. High-quality shape from multi-view stereo and shading under general illumination. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 969–976. 34

[161] Xu, L., Ren, J. S., Liu, C., and Jia, J. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems* (2014), pp. 1790–1798. 87

[162] Yu, T., Wang, H., Ahuja, N., and Chen, W.-C. Sparse lumigraph relighting by illumination and reflectance estimation from multi-view images. In *ACM SIGGRAPH 2006 Sketches* (2006), ACM, p. 175. 67

[163] Yu, Y., Debevec, P., Malik, J., and Hawkins, T. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (1999), ACM Press/Addison-Wesley Publishing Co., pp. 215–224. 67

[164] Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 2528–2535. 109

[165] Zhang, L., et al. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multiview stereo. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 618–625. 35, 36

[166] Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence 21*, 8 (1999), 690–706. 104

[167] Zheng, Z., Lizhuang, M., Zhong, L., and Chen, Z. An extended photometric stereo algorithm for recovering specular object shape and its reflectance properties. *Computer Science and Information Systems 7*, 1 (2010), 1–12. 38, 55

[168] Zhou, T., Krahenbuhl, P., and Efros, A. A. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3469–3477. 87, 104, 105, 107, 110

[169] Zhou, Z., Wu, Z., and Tan, P. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1482–1489. 32, 36, 37, 38, 52, 142

[170] Zhukov, S., Iones, A., and Kronin, G. An ambient light illumination model. In *Rendering Techniques' 98*. Springer, 1998, pp. 45–55. 132

[171] Zickler, T., Ramamoorthi, R., Enrique, S., and Belhumeur, P. N. Reflectance sharing: Predicting appearance from a sparse set of images of a known shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 8 (2006), 1287–1302. 3, 86

Curriculum

Stamatios Georgoulis was born in Thessaloniki, Greece on January 19, 1988. He obtained his diploma in Electrical and Computer Engineering, with honors, from the Aristotle University of Thessaloniki, Greece in March 2011. He conducted the research for his diploma thesis on Adaptive Pain Detection via Webcam using Advanced Image Processing Techniques in collaboration with Dr. Stefanos Eleftheriadis under the supervision of Prof. Leontios Hadjileontiadis. During his stay in Thessaloniki, he also worked as a student research assistant in the Signal Processing and Biomedical Technology Unit, contributed in research papers and technical reports at Signal Processing conferences, participated in Microsoft's Imagine Cup with the research project "Epione: An Innovative Pain Management Solution" for two successive years and represented Greece in the World Finals. In January 2013 he joined the PSI-VISICS research group at the Department of Electrical Engineering (ESAT), KU Leuven, Belgium as a PhD student under the supervision of Prof. Luc Van Gool, working closely with Dr. Marc Proesmans. His research focuses on extracting surface characteristics and lighting in 3D reconstruction from uncalibrated images but he is also interested in solving Computer Vision/Graphics problems using only readily available consumer equipment. In ESAT, he assumed a teaching assistant role for the course 'Digital Electronics and Processors'. During his PhD studies, he has contributed several research papers in high-tier international conferences and journals, such as ICCV and PAMI, and collaborated with exceptional researchers, such as Prof. Luc Van Gool, Prof. Tinne Tuytelaars, Prof. Mario Fritz and Prof. Tobias Ritschel.

Publications

Journal Articles

- **S. Georgoulis**, V. Vanweddingen, M. Proesmans and L. Van Gool, *Shape and Reflectance Using a Camera with Flash.* Submitted in IEEE International Journal on Computer Vision (IJCV).
- **S. Georgoulis**[1], K. Rematas[1], T. Ritschel, E. Gavves, M. Fritz, L. Van Gool and T. Tuytelaars, *Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning.* Published in IEEE Journal on Pattern Analysis and Machine Intelligence (PAMI) 2017.

Conference Articles

- **S. Georgoulis**, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars and L. Van Gool, *What Is Around The Camera.* Published in IEEE International Conference on Computer Vision (ICCV) 2017.
- D. Neven, B. De Brabandere, **S. Georgoulis**, M. Proesmans and L. Van Gool, *Fast Scene Understanding for Autonomous Driving.* Published in IEEE Symposium on Intelligent Vehicles (IV) 2017.
- **S. Georgoulis**[1], V. Vanweddingen[1], M. Proesmans and L. Van Gool, *Material Classification under Natural Illumination Using Reflectance Maps.* Published in IEEE Winter Conference on Applications of Computer Vision (WACV) 2017.
- **S. Georgoulis**, V. Vanweddingen, M. Proesmans and L. Van Gool, *A Gaussian Process Latent Variable Model for BRDF Inference.* Published in IEEE International Conference on Computer Vision (ICCV) 2015.

[1]Denotes equal contribution.

- A. Pevar, L. Verswyvel, **S. Georgoulis**, N. Cornelis, M. Proesmans and L. Van Gool, *Real-time Photometric Stereo.* Published in Photogrammetric Week (PHOWO) 2015.
- **S. Georgoulis**, M. Proesmans and L. Van Gool, *Tackling Shapes and BRDFs Head-on.* Published in IEEE International Conference on 3D Vision (3DV) 2014.

During the MSc studies:

- D. Tzionas, K. Vrenas, S. Eleftheriadis, **S. Georgoulis**, P. C. Petrantonakis and L. J. Hadjileontiadis, *Phantom Limb Pain Management Using Facial Expression Analysis, Biofeedback and Augmented Reality Interfacing.* Published in ACM Software Development for Enhancing Accessibility and Fighting Info-exclusion (DSAI) 2010.
- **S. Georgoulis**, S. Eleftheriadis, D. Tzionas, K. Vrenas, P. Petrantonakis and L. J. Hadjileontiadis, *Epione: An Innovative Pain Management System Using Facial Expression Analysis, Biofeedback and Augmented Reality-Based Distraction.* Published in International Conference on Intelligent Networking and Collaborative Systems (INCoS) 2010.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
PSI
Kasteelpark Arenberg 10 - box 2441
3001 Leuven
sgeorgou@esat.kuleuven.be
http://homes.esat.kuleuven.be/ sgeorgou/